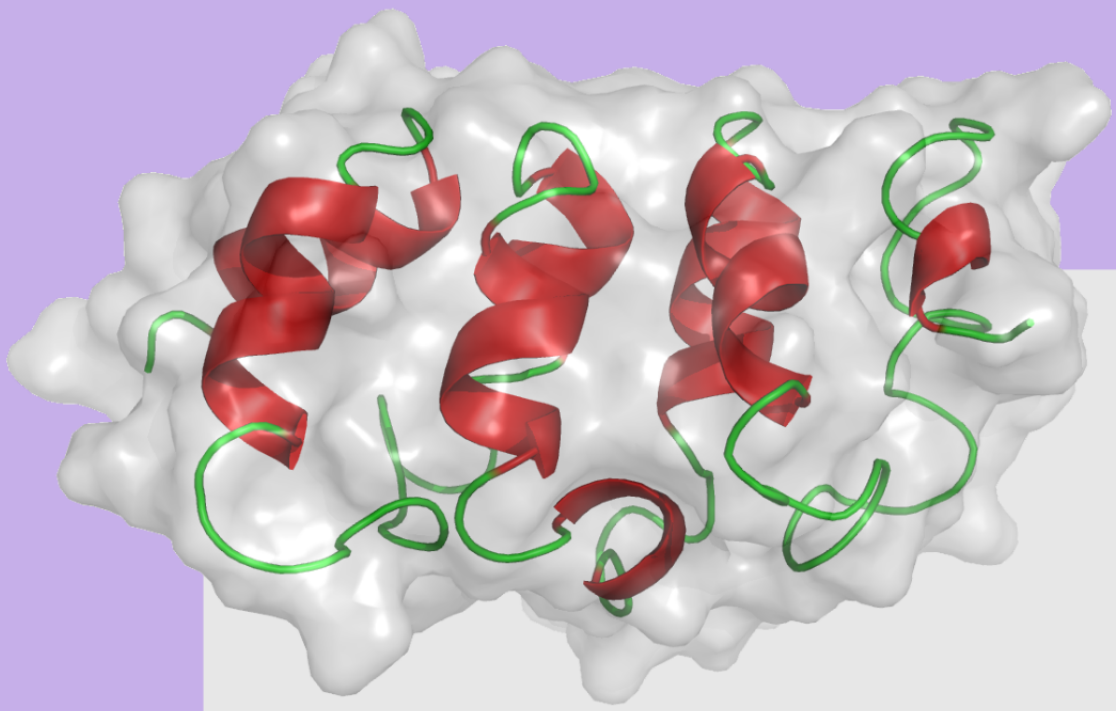
Doctoral Thesis:

Analysis and Development of Algorithms for the Characterization of Biopolymers and Application to Protein Sequencing and Protein Folding

Mauro Faccin



Thesis Adivisor:
Pierpaolo Bruscolini
Zaragoza, October 2011

Departamento de
Física Teórica
Universidad Zaragoza

Analysis and Development of Algorithms for the Characterization of Biopolymers and Application to Protein Sequencing and Protein Folding.

Mauro Faccin

Departamento de Física Teórica
Instituto de Biocomputacón y Física de los Sistemas Complejos
Universidad de Zaragoza

Tesis Dirigida por:
Pierpaolo Bruscolini

Junio, 2011

Universidad Zaragoza

Contents

Resumen

El estudio de esta Tesis se centra en de dos problemas biológicos distintos aplicando un enfoque estadístico-mecánico común: en el primer caso el trabajo se centra en el problema de la secuenciación de los péptidos a través de la interpretación *de novo* de espectros de masa tándem, mientras que en el segundo se estudia el plegamiento de la Miotrofina, una pequeña proteína de tipo modular que ha sido recientemente caracterizada por diferentes técnicas experimentales.

Ambos problemas, como la mayoría de los problemas de tipo biológico, son demasiado complejos como para ser tratados de una manera detallada, *ab initio*, basándose en una descripción microscópica de las interacciones y de los procesos fundamentales involucrados; generalmente para poderlos tratar, se tiene que introducir varias simplificaciones en el modelo del sistema original, a costa de incluir, inevitablemente, ciertas aproximaciones. El típico enfoque empleado en la descripción de sistemas biológicos desde un punto de vista mecánico-estadístico se basa en la traducción efectiva del sistema complejo original a un modelo físico tratable, posiblemente involucrando el uso de coarse-graining u otros tipos de simplificaciones.

En el caso tratado en esta Tesis hemos reescrito el problema original mencionado antes, en términos de un oportuno sistema mecánico estadístico a través de la definición de las variables dinámicas interesantes y de una función energía que determina su comportamiento. Además en ambos

problemas hemos forzado una modelización de las interacciones que nos permita calcular una solución exacta de la distribución de probabilidad en el equilibrio a través del método de la matriz de transferencia.

Ambos problemas están relacionados con proteínas: su identificación con MS/MS y su plegamiento.

Secuenciamiento de Proteínas En la primera parte de la Tesis afrontamos el problema de la secuenciación de péptidos en el marco de la Espectrometría de Masa, que consiste en la interpretación de un espectro de masa para encontrar la secuencia de amino ácidos del péptido estudiado. Gracias a su simplicidad y bajo coste, el uso de esta herramienta se utiliza ampliamente en el campo del análisis bioquímica de muestras desconocidas de proteínas, y, generalmente, está incrustada en una instrumentación de tipo high-throughput automatizada, que produce una gran cantidad de datos, necesitando entonces una herramienta automatizada para interpretarlos.

En principio, un espectro de masa tándem contiene toda la información necesaria para encontrar la secuencia de amino ácidos del péptido que generó el espectro mismo. En la práctica, encontrar dicha secuencia desde un espectro es una tarea muy difícil por la presencia de ruido, picos debidos a iones contaminantes, o falta de fragmentaciones, entre otras cosas. De echo, cada espectro es el resultado de las reglas microscópicas que gobiernan la transferencia de energía y la fragmentación estocástica del péptido precursor en las colisiones, en presencia de ruido y contaminantes de diferentes tipos. Desafortunadamente, la predicción *ab initio* del espectro resultante a partir de la secuencia del péptido precursor es impracticable, si no imposible, y la interpretación de la secuencia del péptido involucra el uso de funciones coste *ad hoc* para medir el acuerdo entre el espectro teórico de un péptido precursor y el espectro experimental. Además, el espacio de búsqueda de las secuencias es generalmente limitada a las secuencias de proteínas ya conocidas ("búsqueda sobre bases de datos"). Gracias a esta restricción la búsqueda de la secuencia resulta más practica y eficiente de los métodos *de*

novo que infieren la secuencia del péptido a partir solo de la información contenida en el espectro, pero está afectada por sus limitaciones. Finalmente en ambos métodos es importante subrayar que un problema central es la asignación de un "grado de confianza" a las predicciones, entre las cuales se pueden encontrar falsos positivos, o sea secuencias equivocadas con alta puntuación.

A continuación describimos un nuevo algoritmo basado en la traducción del problema de la interpretación, al análisis de la distribución al equilibrio de un sistema físico discreto adecuadamente definido, cuyas variables dinámicas describen la presencia de un enlace peptídico (un sitio de fragmentación) en los nodos de un retículo unidimensional, etiquetado por un índice de masa. La función energía *ad hoc* que gobierna el modelo, se deduce de la distribución fenomenológica de los iones y de los picos de ruido en un conjunto de espectros experimentales. Las interacciones que caracterizan esta Hamiltoniana son interacciones locales o a primeros vecinos de manera que la función de partición del modelo puede ser calculada exactamente con el método de la matriz de transferencia. Mientras que la identificación de la secuencia del péptido está asociada a la caracterización del estado fundamental a temperatura cero, la introducción de una temperatura paramétrica y el estudio de las variables termodinámicas como función de la temperatura pueden dar una idea de la calidad de la interpretación, sin apoyarse a bases de datos decoy (bases de datos compuestas por secuencias equivocadas pero con alta puntuación) o en la distribución fenomenológicas de las puntuaciones de los falsos positivos.

Aumentando la temperatura, el equilibrio del sistema se aleja de un régimen dominado por la energía y polarizado hacía la secuencia mejor puntuada, para alcanzar un régimen dominado por la entropía debida a otras secuencias con una menor puntuación, lo cual puede ser útil para evaluar la bondad de la predicción.

El algoritmo ha sido testado sobre un conjunto de espectros experimentales acoplados con la secuencia teórica del péptido precursor y sobre

el mismo conjunto han sido testados algunos de los algoritmos *de novo* disponibles (NovoHMM, PepNovo, Lutefisk). Nuestro algoritmo produce resultados comparables con los programas existentes y además exhibe algunas características útiles relacionadas con el sistema termodinámico asociado que no se encuentran en los demás. Sobresale como característica del algoritmo la posibilidad de controlar la temperatura de trabajo que, junto con la posibilidad de calcular exactamente la distribución de probabilidad al equilibrio, proporciona la oportunidad de considerar al mismo tiempo el entero espacio de las secuencias cada una con su "peso" termodinámico.

Plegamiento de la Proteína Miotrofina En la segunda parte de la Tesis nos hemos centrado en el problema del plegamiento del las proteínas, tratando de caracterizar en particular el equilibrio y la cinética de la Miotrofina, una proteína de tipo modular. El plegamiento de proteínas es un problema que ha sido enormemente estudiado y analizado por parte de los teóricos a partir de muchos niveles de coarse-graining, empezando por las simulaciones a todos los átomos con interacciones realistas, a una variedad de modelos de coarse-graining.

La proteína estudiada en este trabajo, la Miotrofina, es una proteína modular compuesta por cuatro módulos con la misma estructura secundaria aunque diferentes secuencias, dispuestos en una conformación lineal. Esta conformación, común a las proteínas modulares, las diferencia de las ampliamente estudiadas proteínas globulares. En las proteínas globulares las regiones alejadas en la secuencia se acercan y entran en "contacto" cuando la proteína se encuentra en el estado nativo, y el alcance de los contactos puede abarcar la entera molécula. En las proteínas modulares solo hay contactos dentro de cada módulo y entre amino ácidos pertenecientes a módulos contiguos. La Miotrofina muestra un comportamiento interesante que ha atraído la atención de los investigadores: la molécula parece plegarse de manera cooperativa, característica típica de las proteínas globulares, mientras su estructura modular sugiere una independencia intrínseca de cada

módulo, naturalmente asociado a un plegamiento con varios estados intermedios que corresponden al plegamiento independiente de los módulos.

En este trabajo nos basamos en el modelo propuesto por Wako- Saito-Muñoz-Eaton (WSME) para caracterizar el plegamiento de la Miotrofina. Este simple modelo se basa en el uso de una variable binaria para describir el estado ("plegado" o "desnaturalizado") de cada residuo usando la información contenida en la estructura nativa de la molécula, la cual está almacenada en un mapa de contactos que describe la proximidad euclídea de dos residuos y de esta forma predecir el comportamiento en el proceso de plegamiento. A pesar de la presencia de interacciones a largo alcance, la forma especial de la función energía del modelo, le otorga la posibilidad de evaluar exactamente la función de partición, así como las cantidades termodinámicas en el equilibrio, como la energía libre y la capacidad calorífica, mientras que para describir la dinámica del sistema hay que usar unas simulaciones de tipo Monte Carlo.

Usamos el modelo WSME para estudiar la molécula wild-type como algunas mutaciones y por esto tratamos las mutaciones puntuales como variaciones del potencial de interacción del residuo interesado.

En ambos casos, los resultados muestran un buen acuerdo con los datos experimentales, haciendo particular hincapié en la heterogeneidad de los caminos de plegamiento. El minucioso control sobre las simulaciones, que sobrepasa las posibilidades experimentales, nos otorga la posibilidad de sugerir las características de los procesos subyacentes, entre ellos sobresale el acuerdo entre la estructura modular, que produce un perfil de la energía libre con múltiple mínimos, y la cooperación en el proceso de plegamiento, y además sobresale que la simetría entre los caminos de plegamiento y de desnaturalización de la molécula no es una condición necesaria.

Introduction

This Thesis addresses two different biological problems with a common statistical-mechanics approach: in the first case we focus on the problem of peptide sequencing by *de-novo* interpretation of Tandem Mass spectra, while in the second we analyse the folding behaviour of Myotrophin, a small repeat protein that has been recently characterized by different experimental techniques.

Both problems, as the majority of biologically relevant systems and processes, are too complex to be approached in a detailed, *ab-initio* way, based on a reliable, microscopic description of the interactions and fundamental processes involved; on the contrary, typically one has to resort to various simplifications in the modelling of the original system, introducing, as a consequence, some approximations in order to achieve treatability. Hence, the typical approach to the description of a biological system from a statistical-mechanics point of view starts with the research of a successful mapping of the original complex problem to a treatable physical model, possibly involving coarse-graining or other kind of simplifications.

In our case, the mapping of the above mentioned problems to suitable statistical mechanical systems is performed through the definition of the interesting dynamic variables and of an energy function which determines their behaviour. Moreover, for both problems we enforce a modelling of the interactions that is amenable of an exact solution for the equilibrium

probability distribution through the Transfer Matrix Method.

Both problems are related to proteins: their identification with MS/MS and their folding. In the following, we resume some basic facts about such important biomolecules: their synthesis, their structure and their role in cellular processes.

What is a Protein? Proteins are ubiquitous molecules in living cells, involved in almost all the biochemical processes that govern their evolution. Proteins are linear polymeric chains, composed by different combinations of 20 types of amino acids, covalently linked in a sequence (the "primary structure") that completely encodes their three-dimensional structure and function. Each protein is completely characterized by its sequence, whose length may vary between a few tens to several hundred "residues", as the amino acids are called within the protein chain (the name acknowledges the fact that they loose a water molecule upon binding in the main chain).

Protein synthesis is performed inside the cell through a DNA translation mechanism, based on the information contained in the regions of the chromosomes corresponding to the genes. The information on protein composition is, in fact, encoded into the DNA double helix. The latter is found packed in the cell nucleus and is locally unpacked "on demand" and copied to a messenger mRNA strand (DNA transcription). The mRNA strand exits the nucleus to reach the cytoplasm where the ribosomes are located; here, it can undergo some modification (removal of the introns, alternative splicing) before being translated to a polypeptide chain. Ribosomes decode the messenger RNA using the tri-nucleotide translation rules, that assign to each triplet of RNA basis ("codons") a specific residue to be appended to the peptide chain.

Once the amino acid chain is built up, it undergoes the process of folding to reach its functional three-dimensional structure (the "native structure"), which is usually a precisely determined compact structure. Often at this

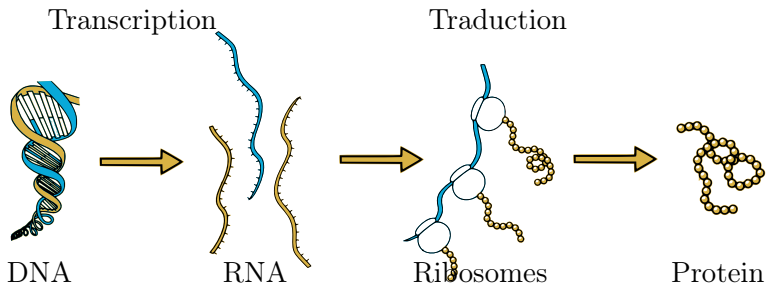
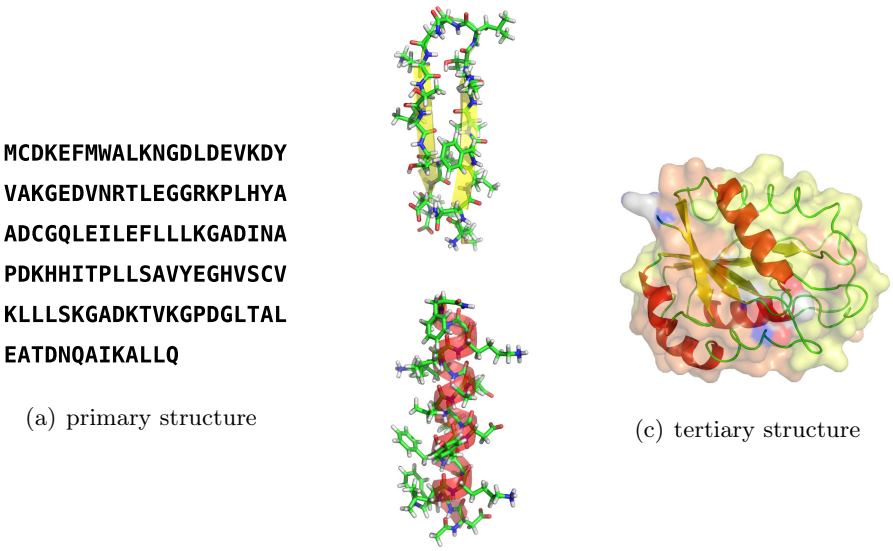


Figure 1: Schematic representation of the DNA transcription and translation processes. The information coded in the DNA is transcripted into an messenger RNA to exit the nucleus where it is stored. Once in the cytoplasm the mRNA is decoded by ribosomes, the latter assign to each nucleotide triplet a specific residue, conforming the final polymeric structure of the protein.

stage the newly formed protein undergoes some modification, mainly characterized by the addiction of functional groups to polypeptide chain. Notice that these changes cannot be read from the genetic code of the DNA, and depend on the specific cell conditions at the moment the protein is synthesized.

Proteins fold in stable structures, dominated by recurring structural motifs, mainly α-helices and β-sheets, called "secondary structures". The local regularity of these motifs is related to the strong directionality of the hydrogen bonds that dictate their geometry. These motifs are arranged in the global tertiary structure of the folded molecule. A quaternary structure is defined by proteins complexes where several proteins are arranged in a functional superstructure. The above "structural paradigm" applies to all proteins with the exception of the Intrinsically Unfolded Proteins, that are functional in a disordered conformation; however, such proteins usually get structured upon binding to their target molecule, so that we can say that the sequence-structure-function relationship is still valid, in a loose sense.



(a) primary structure

(b) secondary structures

(c) tertiary structure

Figure 2: The polymeric chain of a protein is determined by (a) the residue sequence. The resulting chain naturally tends to assume a compact structure mainly conforming (b) structural motifs such as α-helices and β-sheets that are arranged on the final (c) native structure.

Proteins folds naturally in biological conditions, experiencing an out-of-equilibrium process that, starting from an unstructured configuration, ends on the folded native state. On the other hand, folding can be studied in vitro under equilibrium and our of equilibrium conditions: from a thermodynamic point of view, the protein native state, in physiological conditions, represents the stable macro-state, with minimal free-energy; a sudden change to different conditions results in a chemical kinetics that can be usually related to the final free-energy landscape, and is characterized by the search of the most convenient pathway to reach the new equilibrium. In this sense, a folding pathway is specified in terms of the time dependent formation of the long-range contacts between non-neighbouring residues.

The principal folding mechanisms that have been identified can be summarized in four main behaviours [60]: a) Diffusion-Collision: secondary structure folds first and then the motif diffuse and collide to form the correct structure; b) Hydrophobic Collapse: due to the wet environment triggering the folding process, the protein collapses to a molten globule to protect the non-polar residues from the polar solvent, then the search for the right structure is restricted to compact conformations; c) Nucleation-Propagation: a local nucleus starts the folding process and then propagates to the entire structure; d) Nucleation-Condensation: the folding nucleus is composed of non-local contacts, involving distant residues along the chain; once a critical number of such contacts is formed, it triggers the completion of the folding process.

Different proteins, or even the same protein under different conditions, can show different folding mechanisms. The detailed characterization of protein folding pathways passes through the characterization of every conformation that the molecule assumes along the way to the native state[14] and reflects the picture of the energy landscape. Such characterization has been successfully achieved in small proteins combining experimental results with molecular dynamic simulations: most often, the former provide snapshots of the structures along the folding pathway that, although devoid of

the temporal information on contact formation, can give a picture of the intermediate states and the overall folding mechanism for the process, while the latter can give insight on the microscopic details of the relevant structures along the folding pathway.

Protein Sequencing In the first part of the Thesis we focus on the problem of peptide sequencing by Tandem Mass Spectrometry, which consists in the interpretation of a mass spectrum to find the amino-acid sequence of a target peptide. Tandem Mass Spectrometry, thanks to its simplicity and low cost, is vastly used in the field of biochemical analysis of unknown samples of protein, and is generally embedded in an automated high-throughput pipeline, which produces a huge amount of data, requiring an automated interpretation tool.

In principle, a Tandem Mass spectrum contains all the necessary information to find the peptide sequence it comes from. In practice, reading out the sequence from the spectrum is always a difficult task, complicated by the presence of noise, of peaks from contaminants, and of missing or unexpected fragmentations. Actually, each spectrum is the statistical outcome of the microscopic rules governing the energy transfer and stochastic fragmentation of the precursor peptide under collisions, in the presence of contaminants and of noise of different kind. Unfortunately, ab-initio predictions of the spectrum given a precursor ion are impractical, if not impossible, and the identification of the peptide sequence involves the use of ad-hoc score functions to rate the agreement between the theoretical spectrum of a parent peptide and the experimental one. Moreover, the search space is usually limited to the sequences of known proteins ("database search" approach), which is more practical and efficient, but also more limited than de-novo methods which infer the peptide from just the information contained in the spectrum. Finally, it is remarkable that, both for database-search and for de-novo methods, a central problem is related to assessing the reliability of the prediction, mainly related to the prediction of false positives, i.e., wrong

sequences with high scores.

We describe a new algorithm for protein sequencing based on the mapping of the interpretation problem onto the analysis of the equilibrium distribution of a suitably defined discrete physical model, whose dynamic variables describe the presence of a peptide bond (a fragmentation site) on the sites of a one-dimensional lattice, labelled by a mass index. The model is governed by an *ad hoc* energy function, learned from the distribution of the ions and of the noise peaks on a dataset of experimental spectra. Such Hamiltonian is characterized by just on-site and nearest-neighbour interactions, so that the partition function of the model can be exactly calculated by a transfer-matrix approach. While the identification of the precursor peptide is associated to the characterization of the ground-state of the model (the zero-temperature solution), the introduction of a parametric temperature, which represents an original approach to this problem, and the study of the thermodynamic variables as a function of the temperature, can give insight on the quality of the identification, without relying on decoy databases or on phenomenological distribution of the scores of the false positives. Indeed, as the temperature increases, the thermodynamic equilibrium shifts from a regime dominated by the energy and polarized on the best scoring peptide sequence, to a regime dominated by the entropy coming from the alternative sequences with a reasonably good score on the experimental spectrum.

The resulting algorithm has been applied to a test database of spectra and the results are reported in this thesis. We show that the performance of this algorithm is comparable of existing and more popular models, while presents a new valuable feature: the fictitious temperature that pulls the system outside the energetic minimum can give the user a tool to estimate the quality of the prediction.

Protein Folding In the second part of the Thesis we deal with the protein folding problem, focusing in particular on the characterization of the equi-

librium and kinetics of the repeat protein Myotrophin. Protein folding is a challenging problem that has been vastly analysed by theoreticians using different levels of coarse-graining, from all atoms simulations with realistic interactions, to a variety of coarse grained models.

The protein studied in this work, Myotrophin, is a repeat protein composed by four modules with the same secondary structure (but different sequence), arranged in a linear way. This kind of structural organization, common to all repeat proteins, makes them different from the commonly studied globular proteins: in the latter, far apart regions of the sequence can come close ("in contact") to one another in the native structure, and the contact range can span the whole sequence length. On the contrary, in the former, contacts take place just within a module and between contiguous modules. Myotrophin shows a challenging behaviour, that has attracted the attention of the researchers, as the molecule seems to fold in a cooperative way, which is more typical of a globular protein, while its modularity would suggest an intrinsic independence of the conforming modules, more naturally associated to a multi-state folding, with several intermediate states corresponding to the folding of independent modules. Moreover, the experimental results on kinetics have been interpreted in the framework of the existence of two alternative folding pathways, labelled according to which of the two protein ends folds first.

In this work we rely on the Wako- Saito-Muñoz-Eaton (WSME) model to characterize the behaviour of Myotrophin. This simple model resorts to a binary variable to describe the state ("folded" or "unfolded") of each residue, and uses the information on the atomic coordinates of the native structure of a protein, stored in a contact map that describes the euclidean proximity of two residues, to predict its folding behaviour. Despite the presence of long-range interactions in the energy function of the model, its special form allows the exact evaluation of the partition function, so that equilibrium thermodynamic quantities such as free-energies and heat capacities can be exactly and efficiently calculated, while the dynamics can be studied by

Monte Carlo simulations.

We use a slightly modified WSME model to study both the wild type protein and some mutants; to this purpose, we mimic single point mutations by changing the interactions potentials of the mutated residue. We introduce a new way to characterize, in the folding and unfolding processes, the relaxation pathways based on the secondary structure formation and denaturation.

In both the wild type and mutant proteins, results show a good qualitative agreement with experimental results, with particular emphasis on the heterogeneity of the folding pathways. The higher control allowed by *in silico* simulations over *in vitro* experiments, and an original way to interpret kinetic data, let us shed some light on the subtending processes as the consistency of the modular structure and free-energy profile and the folding cooperativity, as well as the lack of symmetry between the folding and unfolding pathways.

Part I

De Novo Interpretation of Tandem Mass Spectra

Introduction to the Problem

The first part of the thesis deals with the application of statistical mechanics methods to the interpretation of the experimental spectra obtained with the Tandem Mass Spectrometry technique (usually abbreviated as MS/MS, or MS2). The interest for the subject is two-fold: from a purely theoretical point of view, we will show that the problem of the interpretation of noisy MS/MS spectra can be mapped on the search for the equilibrium distribution of the states of a suitably-designed one-dimensional lattice system. The latter can be solved exactly, by a transfer-matrix technique, under some approximations that we will discuss in Chapters 2. On the other hand, MS/MS spectra interpretation is nowadays a key step in high-throughput peptide and protein sequencing, a central subject in many proteomics studies, with important implications also for genomics and metabolomics. To provide the reader with a better understanding of the relevance of the experimental technique, in the following we give an overview of the application of MS/MS to proteomics.

Even if several definitions have been proposed, we can simply say that proteomics is the study of the proteome of an organism: that is, the identification of all possible protein contents of a cell, and their characterization in terms of quantity, structure, function, and molecular interactions.

Proteomics arises as the "natural" counterpart, in the protein domain, of genomics, that studies the genetic information of an organism. Indeed, sometimes the proteome has been simply identified as the protein expression of the genome[27, 79]. This view is due to the fact that the entire information, characterizing the biochemical processes of any living system, is supposed to lay inside the genetic code; moreover, it is also fostered by the enormous growth of genetic data that have been collected in recent years, with the complete decoding of the genomes of several organisms.

However, the above view is somewhat restrictive and incomplete, since it suggests that protein content is univocally defined from the genetic code,

and a direct map can be performed between genome and proteome. In reality the information coded in the DNA undergoes several levels of modifications until it reaches the final state of usable protein. During DNA transcription, information coded into DNA is copied to the mRNA, which, in eukaryotes, can undergo some post-transcriptional modifications and intron splicing. Moreover, the protein, obtained upon the translation of mRNA often undergoes some post-translational modifications, mainly consisting in the addition of cofactors or small chemical groups, which are related to its function.

This transcription-translation-modification process, in turn, is controlled by the type and amount of the proteins and signalling molecules that are present in the cell, providing a complex feedback mechanism that regulates any cell activity. As a result, the proteome of any organism is highly dynamic and changes over time; the types of expressed proteins, their abundance, state of modification, sub-cellular location, etc. depend on the physiological state of the cell, on the cell location and on the tissue the cell belongs to, thus modifying and giving a inherent and time-dependent structure to the static, overall information that can be obtained by simply translating the genes into protein sequences.

Needless to say, accessing the dynamical information on the changing protein content of a cell is even more important that the static knowledge of its genetic information, to shed light on how a cell functions at the molecular level, how it reacts to adjust to external conditions, how we can distinguish a tumor cell from a normal one, just to mention a few possible questions.

For the above reasons, a fundamental goal in proteomics is represented by the complete analysis of the protein content of a given sample, and answering a number of qualitative and quantitative questions about the cell-sample content. Such analysis begins with the identification of the proteins expressed in a cell, together with the post-translational modifications they carry: the latter can provide important information, that cannot be read from the genome, on the biological processes in which the protein is involved.

Several methods can be used to identify proteins, depending on the degree of previous knowledge available: if the question is to find out if a known protein is present in a mixture of a few proteins, low resolution techniques can be sufficient, consisting in a purification step to isolate the protein of interest from the rest (for instance, by gel electrophoresis or mass spectrometry), plus an identification test probing its interaction with a suitable substrate (for instance, using antibody probes, as in Western Blotting or ELISA tests).

Another intermediate resolution technique is protein fingerprinting, consisting in cutting the protein of interest, proceeding from of a certain organism, into smaller peptides, with some specific enzyme acting in a known way (for instance, cleaving at every occurrence of a particular amino-acid), and measuring the mass of the resulting peptides. Identification is then performed by matching the protein and peptide masses to those obtained by considering the database of all the proteins of that organism, simulating the action of the enzyme on each of them, an calculating the mass of the resulting peptides. This technique is very efficient provided that the original sample contained a well purified protein and not a mixture, and, obviously, if the database of all the possible proteins, usually derived from the knowledge of the genome of the organism, is available.

However, at the bottom line, if no previous knowledge on the possible sequences is available, direct sequencing is the only possibility to identify the primary structure of a protein. Until the decade of the nineties, Edman degradation was the only available technique to that goal. In this approach, amino-acid residues are chemically removed and identified one by one, starting from the N-term of the protein. This technique is very precise, but also too slow to be applied in high-throughput proteomics. On the other hand, Tandem Mass Spectrometry allows the analysis of a protein in a few seconds, it is able to detect post translational modification, and is therefore the technique of choice in high-throughput proteomics. However, its accuracy is often jeopardized by the fact that the resulting spectra can be quite noisy,

difficulting *de-novo* and also database assisted interpretation.

In the next chapters we will discuss in more details the experimental technique, the present interpretation strategies, and our proposal for a *de-novo* interpretation, not relying on a protein database.

In Chap. 1 we will introduce shortly the technique of Tandem Mass Spectrometry (MS/MS) and the instrumentation used to identify proteins and mixtures of proteins. Namely, we will describe the experimental workflow based on proteins purification, cleavage and separation, that yields the ionized peptides, called precursors, that are then fragmented and recollected on the experimental spectrum, ordered on the basis of the mass-to-charge ratio.

The resulting huge number of experimental spectra must be interpreted independently, generally by the use of an external software. In the second part of Chap. 1 we will describe the different types of software, focusing on the difference between database-based sequencing tools (the most popular), and *de-novo* sequencing algorithms.

Then we will introduce a new approach to the problem of data interpretation. In particular, in Chap. 2, the identification of the residue sequence, underlying a given spectrum, is turned into the study of the equilibrium of a unidimensional statistical mechanics system, with an energy function containing the experimental spectrum as an external field. The search of the amino acid sequence that most likely produced the target spectrum, can then be approached with the classical tools of statistical mechanics, by looking for the equilibrium distribution of the system; the solution for the latter at low temperature can give us an insight in the most probable peptide sequence. Moreover, the study of the equilibrium state at higher temperature informs on the energy landscape induced by the target spectrum on the sequence space; this allows the introduction of a quality test for the prediction.

We will discuss how the equilibrium state can be exactly calculated, and

which are the relevant thermodynamic observables that allow us to identify the sequence of the precursor and to evaluate the quality of the prediction, when possible.

In chapters 3 and 4 we will define and discuss a suitable cost function to map the sequencing problem to the unidimensional statistical system described above. The cost function will be based on the phenomenological distributions of the spectrum peaks calculated from a known and reliable database.

We will discuss in Chap. 3 the selection, filtering and processing of some freely available spectral databases, to characterize the phenomenological distribution of the peaks corresponding to the different possible fragments. In Chap. 4 we will discuss the definition of an empirical energy function based on the results of Chap. 3.

Finally, Chap. 5 is dedicated to the presentation of the results of the application of our method to a test database of spectra, and to the comparison of its performance with that of other *de-novo* methods, outlining the possible direction for future developments.

Chapter 1

Protein Sequencing by MS/MS

Tandem mass spectrometry (MS/MS) is used in order to produce structural information about a compound by fragmenting-specific sample ions inside the mass spectrometer and identifying the resulting fragment ions. This information can then be pieced together to generate structural information regarding the intact molecule. Tandem mass spectrometry also enables specific compounds to be detected in complex mixtures on account of their specific and characteristic fragmentation patterns. In proteomics, MS/MS represents a useful instrument to identify the primary structure (i.e., the amino-acid sequence) of the proteins contained in the target sample.

A tandem mass spectrometer is a mass spectrometer that has more than one analyser, in practice usually two. The first mass analyser (MS_1) is used to select user-specified sample ions (*precursor ions*) arising from a particular component; usually ions associated to the whole molecule (i.e. $(M+H)^+$ or $(M-H)^-$ ions). The selected ions pass into the collision cell, where they are bombarded by the inert gas molecules which cause fragment ions to be formed (*product ions*). The latter are analysed, i.e. separated according to their mass to charge ratios, by the second analyser (MS_2). The resulting product ions arise directly from the precursor ions specified in the exper-

iment, and thus produce a fingerprint pattern specific to the compound under investigation. This type of experiment is particularly useful for providing structural information concerning small organic molecules and for generating peptide sequence information. In the latter case, which is the one of interest here, the precursors ions are those obtained by enzymatic digestion of an unidentified protein, which typically yields small polypeptides of length ranging between a few and a few tens of amino-acids. Their separation according to their mass, and successive fragmentation into all possible pairs of N- and C-term pieces, provides, in principle, the full information to enable the readout of the sequence from the resulting fragment spectrum, as explained in the following sections.

1.1 The Sequencing Workflow

1.1.1 Separation and Digestion

In a typical proteomics experiment a sample with a complex mixture of protein (for instance, resulting from a cell lysate, and providing a snapshot of the proteome of the cell at the moment of its disruption) has to be analysed. The first step for identification and quantitation of the different proteins contained in the mixture is to separate them from each other, to allow individual processing.

The separation of the sample mixture is usually performed by 2-D gel electrophoresis, producing numerous samples in which there is usually only one dominant protein, even if other molecules and protein can be present at very low concentration. In other words, each spot can be considered as an ensemble of identical protein, if we neglect the above mentioned impurities. Once separated, the proteins undergo a enzymatic digestion usually accomplished by Trypsin.

Trypsin is an enzyme with a catalytic pocket that specifically cleaves the

polypeptide on the carboxyl side of the amino-acids Lysine (K) or Arginine (R). The cleavage is inhibited if the Lysine or the Arginine are followed by a Proline (P). Although other enzymes can be used to cut the protein into smaller peptides, the above rules make trypsin the most useful and common enzyme for digesting proteins: indeed, due to the natural frequency of the K and R amino acids in protein sequences, the above cleaving rules produce a specific length distribution for the tryptic cleavage, with only 10% of fragment longer than 20 residues, while keeping a reasonable low concentration of short residues.

1.1.2 Ionization

Mass Spectrometry relies on the availability of simple ionizations techniques and accurate measurement instrumentation. The most popular ionisation methods available (see [5] and references cited therein) are Electrospray Ionization (ESI) and Matrix Assisted Laser Desorption Ionization (MALDI). Other methods include Atmospheric Pressure Chemical Ionisation, Chemical Ionisation, Electron Impact, Fast Atom Bombardment, Field Desorption and Field Ionisation, Thermospray Ionisation.

Indeed, the development of Electrospray Ionization [16, 19] and Matrix-assisted Laser Desorption Ionization [26] prompted for important developments in mass spectroscopy instrumentation that would revolutionize protein chemistry and protein analysis during the decade of the nineties.

The above methods generate ions from large, non-volatile analytes such as proteins and peptides without significant analyte fragmentation. For this ionization characteristic they are also referred to as "soft" ionization methods. (in fact they are so soft that under specific conditions even non-covalent interactions may be maintained during the ionization process.)

Due to the pre-eminence of this two methods, we discuss them in more detail in the following.

Electrospray Ionization. ESI popularity raised in part due to the ease with which it could be interfaced with popular chromatographic and electrophoretic liquid-phase separation techniques necessary to produce almost pure protein samples. This ionization method, then, displaced previous popular methods such as fast atom bombardment for protein and peptide samples dissolved in a liquid phase.

During standard ESI [80], the sample is dissolved in a polar, volatile solvent and pumped through a narrow capillary. A high voltage is applied to the tip of the capillary where, as a consequence of this strong electric field, the sample emerge and disperse in form of the aerosol of highly charged droplets. A gas flow, usually nitrogen, helps to direct the spray emerging from the capillary tip towards the mass spectrometer. The nitrogen warm flow, also known as drying gas, helps the evaporation of the solvent causing the droplets to diminish in size. The charged ions trapped in the decreasing-in-size droplets, manage to escape and pass through a sampling cone into an intermediate vacuum region, and from there, through a small hole, into the first analyser of the mass spectrometer, which is kept in high vacuum conditions.

The ionization of proteins and peptides is usually positive while for saccharides and oligonucleotides ionization is negative. In all cases, the m/z scale must be calibrated by analysing a standard sample, similar to the target sample being analysed, and then a mass correction must by applied.

The sample ions usually are produced in a singly charged state that in positive ionization, the one we are interested in, means protonated molecular ions $(M+H)^+$ (while negative ionization give rise to deprotonated molecular ions). High values of peptide mass, exceeding 1200 Da., can give rise to multiply charged ions $(M + nH)^{n+}$[70].

MALDI. MALDI is the other popular ionization method, although for different reasons. This ionization method is commonly used coupled to

the time-of-flight (TOF) analyser, that we will talk about later on, that is robust, simple, and sensitive, with a large mass range. MALDI ionization method produce ions which generally are singly charged, simplifying, thus, the interpretation task.

MALDI [25] deals well with proteins and other thermolabile, non-volatile organic compounds, especially those of high molecular mass. The mass accuracy depends on the type and performance of the analyser of the mass spectrometer, but most modern instruments should be capable of measuring masses to within 0.01% of the molecular mass of the sample. In MALDI, the sample is co-crystallised with an highly absorbing organic matrix compound, this usually contain an aromatic ring that can absorb the wavelength of a laser. The laser bombardment is transformed into energy by the matrix compound and then particles of the mixture of matrix compound and analyte ions are released from the surface.

The calibration of the spectrometer scale is performed, usually, with a known sample that can either be analysed independently (external calibration) or pre-mixed with the sample and matrix (internal calibration). MALDI is also a "soft" ionisation method and the predominant products are protonated molecular ions ($M+H^+$) regardless of the molecular mass. Traces of doubly charged ions at approximately half the m/z value can be found.

1.1.3 Analysers

After the ionization of the sample molecular ions, these have to be separated and their mass-to-charge ration measured. A variety of analysers are used to this scope, the most commons are are Quadrupole mass analyser, Magnetic sector and Time Of Flight (TOF).

Quadrupole Mass Analyser. This analyser consist of four parallel rods. Opposing rods are paired together and connected electrically: a variable potential is applied between the two pair of rods with radio frequency oscillations, while a time-independent potential is superimposed. The injected ions are affected by the oscillating electrical potential and the beam is alternatively focused toward the centre and dispersed toward the rods in both directions perpendicular to the rods axis. The superimposed constant potential, say positively biasing the rod pair 1, and negatively biasing the rod pair 2, acts as a mass filter in the two dimension: if positively biased, the heavier ions (suppose positive ions) are affected by the average potential and focused toward the centre of the quadrupole, lighter ions, affected by the rapidly varying potential, experience large accelerations toward the rods and are eliminated from the beam, creating a high pass mass filter; on the other axis, with negative biased potential, the heavier ions are attracted by the average negative rods potential, while in lighter ions this behaviour is contrasted by the oscillating potential, this a low pass mass filter. The combination of the two filters result in a band pass mass filter that select ions based on their mass-to-charge ratio. The band width is related to the instrument mass resolution[49].

Magnetic Sector The behaviour of ions in a homogeneous, linear, static magnetic field as in a sector instrument is simple. The physics are described by the Lorentz force law in the case of zero electric field:

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B} \tag{1.1}$$

where $\mathbf{B}$ is the magnetic field induction, q is the charge of the particle, $\mathbf{v}$ is its current velocity. In this case the molecular ions are accelerated by a electrical potential V and then injected into the magnetic field, the forces experienced by the particles are perpendicular to the trajectory resulting in a arc of square radius $r^2 = \frac{2V}{B^2}\frac{m}{z}$. Given then the accelerating potential V and the magnetic field inside the magnetic sector B, then the radius of

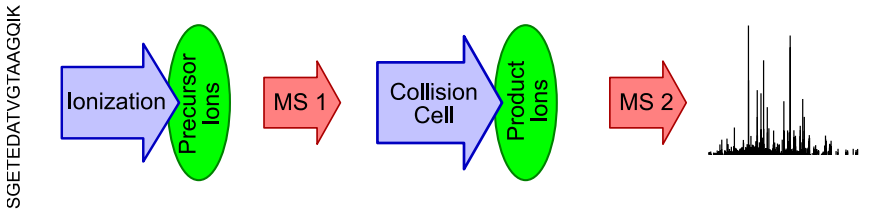


Figure 1.1: MS/MS scheme. The protein, once purified and digested by Trypsin, undergoes ionization through MALDI or ESI, the resulting precursor ions are separated in the first analyser MS_1. A selected precursor ions is fragmented in the CID (see text) and the resulting product ions are separated by the second analyser MS_2 to be recollected in the experimental spectrum on the basis of their mass-to-charge ratio.

the trajectory only depends on the mass-to-charge ratio of the travelling particle.

Time Of Flight Time-of-flight mass spectrometry (TOF) is a method of mass spectrometry in which ions $\frac{m}{z}$ ratio is determined via a time measurement. Ions are accelerated by an electric field $\mathbf{E}$ of known strength:

$$\mathbf{F} = q\mathbf{E} \tag{1.2}$$

this results in an ion acceleration and a velocity that depends only on the mass-to-charge ratio. The time that it takes to reach the detector at a known distance is measured and will depend on the m/z ratio of the particle.

1.1.4 Collision-Induced Dissociation and Product Detection

So far, the peptides, produced from the enzymatic digestion of an ensemble of identical proteins, have been separated according to their mass/charge ratio in the first MS analyser. The determination of their masses can be sufficient for the identification of the protein by peptide fingerprinting, as

described in the previous chapter, provided that a protein database containing its sequence is available, and the identification is not hindered by post-translational modifications that change the mass of the "bare" peptide. However, to proceed further towards a true sequencing of each peptide, a mechanism is needed to resolve the individual amino-acid residues that compose the peptide. To this end, Collision Induced Dissociation is combined to a second MS analyser, according to the following protocol (see Fig. 1.1 for a scheme of the process)..

The molecular ions that travel from the first analyser are accelerated by some electrical potential to high kinetic energy in the vacuum of the mass spectrometer and then allowed to collide with neutral gas molecules (often helium, nitrogen or argon). In the collision some of the kinetic energy is converted into internal energy which results in bond breakage and the fragmentation of the molecular ion into smaller fragments. whose number depends on the energy involved in the collision. Typically, low-energy CID (the one that we will deal with in more detail) split the "parent peptide" in two fragments, a N-terminal and a C-terminal one, while high-energy CID can yield internal fragments, braking the parent peptide in more than one point.

While CID is currently the most popular method for standard tandem mass spectrometry, there are other also other fragmentation methods for special purposes, for example electron-transfer dissociation (ETD), electron-capture dissociation (ECD), and infra-red multi-photon dissociation (IRMPD). Alternative fragmentation methods are particularly useful for identification of phosphorylation sites and other post-translational modifications (PTMs).

Peptides fragment in a reasonably well-documented manner [34, 64] even if the underlying physics is not completely understood, and it is not feasible, at the moment, to predict the outcoming distribution of charged and neutral fragments, since they depend not only on the composition but also on the conformational state of the peptide during the collision[15, 17, 77, 87].

Fragmentation generally occurs between two consecutive residues, breaking a peptide bond on the main chain of the peptide, this can happen in three different ways generating three corresponding pairs of fragments (see Fig. 1.2). The first fragmentation site is localised on the bond between the C_α carbon and the subsequent $C = O$ of the same residue while the proton carrying the positive charge could be alternatively associated to the N-terminal fragment, giving rise to an a type ion, or to the C-terminal fragment, giving rise to a x type ion. The second fragmentation site is between the C-terminal of the carboxyl group of a residue and the amino group of the following residue, giving rise to the b and y type ions for the N and C-terminal parts respectively. Last fragmentation site can be found between the amino group and the C_α carbon of the same residue, giving rise to c and z ions. [35, 64] Only the fragments carrying the charge inherited from the precursor ion produce an event on the resulting spectrum. While the resulting distribution of ions is not predictable, ions labelled b and y are the most observed.

To make things more difficult, during collisions the fragment ions can undergo neutral losses (loosing, for instance, water or ammonia groups) or accept more charges depending on the precursor charge state depending on the residues composing the resulting ion. If the precursor ion is carrying more than one charge, the outcoming singly charged ions can accept a further proton increasing its charge state if the ion contain at least a Lysine (K), an Arginine (R) or a Histidine (H), this results in a observed peak at half the joint mass of the product ion and the proton. Some residues, during the collision dissociation can loose a neutral group, the most common neutral groups are water (lost by Serine (S) and Threonine (T)) that results in a second peak shifted by -18.01 Da. from the original peak, ammonia (Glutamine (G), K, R) with a peak shift of -17.03 Da., and urea (R) with a peak shift of -97.98; while H and R can accept a water group resulting in a new peak shifted by +18.01 Da..

Another difficulty source for a straight spectrum interpretation is the

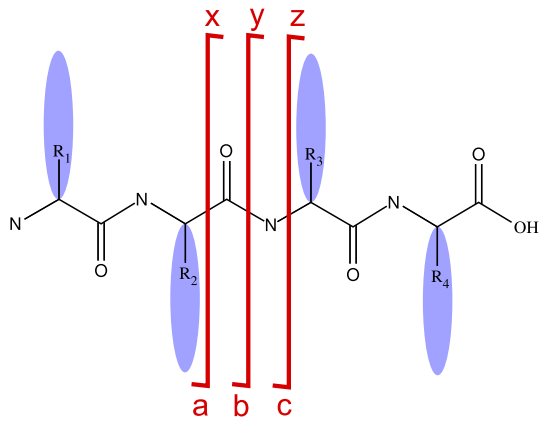
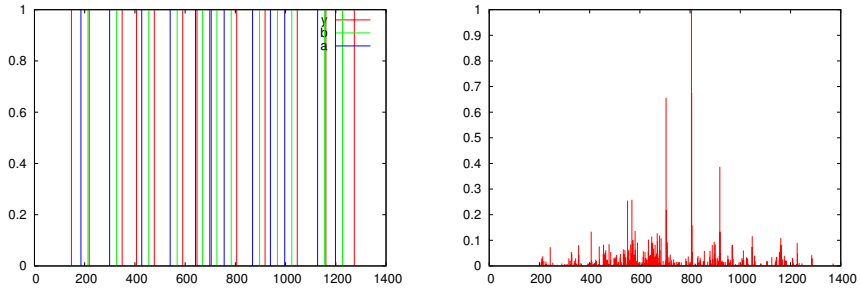


Figure 1.2: Peptide fragmentation scheme. Ions families produced by the CID fragmentation of the precursor ions, distinguished by fragmentation pattern. Fragments a, b and c are the N-term part of the fragmentation in the NH-CH, CH-CO, and CO-NH bonds respectively, while x, y and z correspond to the C-terminal part of the same fragmentations.

presence in the observed ions of elements with high isotopes content, such as carbon and nitrogen, the highest peaks are then often accompanied by isotopes peaks shifted by 1 Da. that become important for bigger ion masses. It is feasible that more than one isotope peak is found if heavier ions are presents.

The resulting charged fragments are detected by the second MS analyser, and add up to the parent peptide spectrum. If fragmentations at every position along the sequence were equally likely, and if no noise from impurities or contaminants were present, the resulting spectrum would consist of a series of N-series and C-series peaks, of equal intensity, as in Fig. 1.3(a). The determination of the sequence would be possible by simply reading out the masses of the component residues from the difference in the position of the peaks along the $m/zaxis$. However, things are more complicated than that, and a typical spectra looks like the one in Fig. 1.3(b), where true peaks



(a) (b)

Figure 1.3: Comparison of the expected theoretical spectrum (left) including only the most observed fragment, y ions (red), b ions (green) ans a ions (blue), with the experimental spectrum (right) for the peptide VINQLTGGLAGMAK.

are mixed with noise peaks, of similar intensity, or can be missing from the spectrum, complicating the task of a correct identification of the underlying parent peptide sequence. Moreover, a typical spectroscopy experiment may produce several hundreds of thousands of precursor spectra that have to be interpreted separately, which calls for an automated interpretation resorting to computer programs.

The final outcoming spectrum is a list of paired values of mass-to-charge ratios and peak intensities where, the first value represent the mass centroid of the observed peak while the second correspond to the peak area or intensity, that is the total number of observed events. Notice that the discrete nature of the resulting peaks is, actually, a product of the automatic post-processing implemented by the analyser software: both for the parent peptide and for the fragments, the raw signal is a continuous one, with a noise background characterized by small fluctuations around a mean intensity, which add to high peaks, more or less intense or well resolved. Not all the peaks can be considered as representing the true signal: they could also come from an extremal noise fluctuation, or from common contaminants of

several kinds. The raw signal is processed to eliminate everything below an arbitrarily chosen value (considered as background noise), and to report in the output file, for each peak, its position (as the peak or the centroid position) and intensity (as the peak height, or sometimes the peak area).

The files provided by the SEQUEST algorithm (*.dta) from Finnigan LCQ data represent one of the most common file format and contain this information. The first line of the file presents the mass of the single charged precursor ion $m(MH^+)$ and its actual charge state Q. From the second line until the end, the file include the list of peaks (ρ_α, I_α).

1.2 Protein Sequencing

Protein sequencing experiments produce an huge amount of MS/MS spectra, and the latter are interpreted in order to infer the precursors composition and sequence. The spectrum interpretation in the ideal case should be solved by simple differences of the m/z ratio of the spectrum ideal peaks, and the conversion of those values in a sequence of amino-acids. This ideal situation is far from the real case which is affected by instrument noise and other complexities intrinsic to the system like the presence of different isotopes in each ions that results in multiple peaks or the internal fragmentations and neutral losses in product ions, as discussed in the previous section. Thus, in the real case spectrum interpretation is managed by essentially two types of computer algorithms: *data-base* search and *de-novo* search. The former is the most popular and actually the most effective and fastest, provided that some conditions are fulfilled. It consists in searching the precursor spectrum against a data-base containing well known sequences. *De-novo* identification relies only on the information enclosed in the spectrum to infer its amino-acid sequence.

1.2.1 Database Sequencing Algorithms

With the increase of popularity of Tandem Mass Spectral peptide sequencing, computer automated algorithms for spectra interpretation become available. They solved the sequencing problem by providing a database of theoretical spectra of candidate peptides.

The most popular protein sequencing algorithms are usually shipped with the spectroscopy instrumentation or maybe purchased separately. Two commonly used algorithms are Mascot[18] and SEQUEST[62], which are closed-source software. They typically correlate the theoretical peptide sequence to the unknown mass spectrum and match them with a correlation or cost function.

SEQUEST, as an example, provide a huge database of known sequenced proteins. After filtering the entire database on the bases of the precursor mass and digestion rules, the program acts in three steps: in the first step, the experimental data are reduced filtering out all but the 200 highest peaks and normalizing them to an intensity of 100, in the second step the algorithm calculates the masses of the expected y and b ions from the database sequences, building a theoretical spectrum, and compares them to the filtered target spectrum through an empiric score function. In the last step the 500 best sequences are then compared by cross-correlation to the experimental fragments, this include the reconstruction of a more accurate theoretical spectrum incorporating less abundant ions and intensity informations based on empirical observations. Once filtered out the precursor ion from the target spectrum and performed a local normalization of the observed ions, a cross-correlation function (XCorr) is calculated to attribute the final score. The difference in score between the first- and second-ranked sequence (ΔCn) is used to distinguish false positives.

Mascot scoring algorithm calculates the probability that the match between experimental data and the theoretical spectrum is a chance event, where the sequence with lower probability is the best candidate. However

the details of the model have not been published.

This kind of approach, based on a database searching algorithm, is very effective in standard condition, but is affected by a number of weaknesses. The use of a sequence database, which is finite per definition, translate in a finite number of possibly sequencing precursors. If the proteome of the cell is absent from the database, the resulting sequence inferences are likely to be incorrect. The target peptide can be absent from the database for a number of reasons, for instance, depending on the purposes of the experiments, the target protein can be a mutated one or can present a post-translational modification (PTMs).

1.2.2 *de-novo* Sequencing Algorithms

Another approach to the peptide sequencing problem is through a *de-novo* algorithm. This new approach selects an amino-acid sequence over some candidate peptides according to a given scoring function. Usually a theoretical spectrum is generated from the peptide, following some known fragmentation rules, and compared to the experimental one. The scoring function, on which each algorithm relies, measures the similarity of those two spectra and finally the best candidate is reported. The scoring function is then the core of the algorithm and the overall quality of the predictions depend on the ability of the scientist to describe all the significant processes the system undergoes.

In this situation the reconstruction of the peptide sequence in not affected by any database constraint. This approach would therefore be very useful in situations where the genome of the organism is unknown or just partially known, providing an incomplete protein database, but also in the case of known genomes, when the existence of some post-translational modifications (PTMs) to specific residues, provokes peak displacements and a combinatorial number of possible alternative interpretations, impairing the *database search* approach for sequencing.

Various algorithms have been developed to solve the *de-novo* peptide sequencing problem. They use different approaches, and various scoring function have been written to select the best precursor peptide sequence.

Lutefisk [72, 73] creates a weighted graph using a simple scoring scheme. A small set of pre-filtered ions from the target spectrum are interpreted alternatively as b and y fragment as it is impossible *a priori* to distinguish between them. Successively b ions and b-type images of y ions are recollected in a "sequence graph". The resulting sequences are build "jumping" from a node to another if separated by a length equal to a residue mass (or multiple residues masses) and sequences connecting N- to C-terminal are stored for further analysis. The sequences set undergoes a series of filters and best ones are selected accounting for the interpretable fraction of peaks intensity.

PEAKS [43] creates a similar weighted graph and at each node an empirical cost-function is evaluated in terms of presence-absence of the relatives y and b ions, increasing their reward if other ions types (like $x, y-NH_3$ and $y-H_2O$ for the C-terminal or $a, c, b-NH_3$ and $b-H_2O$ for the N-terminal) can be matched on the spectrum. A set of the best 10000 sequences are efficiently calculated in PEAKS and in a following step a similar but more stringent and computationally inefficient scoring scheme is applied in order to produce the final output.

Sherenga [15], starting from the target spectrum, constructs a graph in which each node corresponds to a N-terminal peptide fragment, having interpreted a experimental peak as a given ion type (i.e. b type ions). Nodes are connected if they differ of about the mass of a residue and are weighted depending on a learned function of the relative intensities of the peaks produced by the same fragmentation site. The resulting problem is to find the path with the highest score which is solved with a fast dynamical programming algorithm.

PepNovo [23] rely on a probabilistic description of the CID process and generates a network fragmentation model to find the most probable sequence. Similar to the previous model and in some sense an improvement of it, PepNovo compares at each sequence site m the score of a fragmentation site interpretation, based on a semi-empirical probabilistic network of fragmentation rules, with the hypothesis of random events. A weighted graph of fragmentation sites nodes is then constructed, with links connecting nodes separated by a residue mass, and through a dynamical programming algorithm the highest scoring path is found.

NovoHMM [22] use a Hidden Markov Model to approach the peptide sequencing problem. The algorithm combine a learned "transition probability" between two residues and the "emission probability" of the expected spectrum ions. These probabilities are then combined in a factorial Hidden Markov Model. To simplify the model, only y and b ions are modelled and, thanks to the symmetry of the resulting algorithm, sequence is folded and the whole chain divided in two sub-chains. The algorithm start generating sequences at both ends of the chain and the two part are merged when it reach the centre.

However, the *de novo* sequencing problem in tandem mass spectrometry remain an open problem. The information included into mass spectrometry is partially ambiguous, as it is not possible, for example, to distinguish between Leucine L and Isoleucine I relying only on peaks information, since the two residues have exactly the same atomic composition and, hence, mass. Moreover, spectra are usually incomplete and noisy: low values of mass-to-charge are often not observed, while some fragmentation sites are naturally protected from low energy fragmentation like in CID as in Proline bond. Ions other than y and b are usually found with low abundance and often are completely absent from the spectrum. Noise peaks help misinterpretation, they can depends on sensor noise or on sample contamination due to the preparation of the sample for the measure.

Nevertheless the development of algorithms for *de novo* sequencing of product peptides can enhance the biochemical studies expanding the researches possibilities or improving already existing database searching algorithms.

1.2.3 Reliability of the identification

A common problem for both database and *de-novo* approaches is to assess the goodness of the interpretation. A good score, in either database-search of *de-novo* sequencing, does not necessarily imply that the identification is correct. This important issue is somewhat mitigated, especially for database searching algorithms, by the fact that the usual goal in proteomics is to identify proteins, and not individual peptide sequences. In this sense, redundancy can balance the uncertainty: the identification of a certain number of peptides as belonging to a certain protein can be considered as a proof of the correct identification, even if none of the individual interpretations is reliable enough. However, this has some important drawback: in this way it is easy to correctly recognize common and highly expressed proteins, that in some sense could be the "trivial" ones, while important proteins appearing in low abundance can go unnoticed, since their identification is based on the correct characterization of perhaps just one or a few, possibly low quality spectra.

So, the assessment of the reliability of peptide identification is indeed an important issue, and, unfortunately, there is no well-founded theoretical estimate of the significance of the scores, so that different empirical methods have been proposed. For database search, where the sequence space to be searched is much smaller than for the *de novo* case, the approach to this problems involves the definition of a few indicators of the quality of the prediction, as for instance, the p-value of the score, or the gap in the score between the first and second hit in the sequence database. For instance, MASCOT score is itself related to a p-value, since it is based on

the probability that the top hit is a random event, and on the size of the sequence database searched [48]. Since the scores are not really rooted on a sound modelling of the fragmentation process and of the noise distribution, the probability distribution of the score values cannot be derived from first principles, and involve strong approximations. Indeed, it is estimated that just 20% of MS/MS spectra is successfully identified by database-matching algorithms[45].

The problem of the reliability of the identifications is a very well known problem, and several methods have been proposed to assess the value of the predictions, in a post-processing of the sequencing process [39]. The lack of good theoretical estimates for the statistical significance of peptide identification prompts for the use of empirical database-dependent estimates of error rates (resorting to Hypergeometric, Gaussian, Poisson, or other distributions to predict the probability that a fragment ion matches a peak, and to calculate the probability that a match between a peptide and a spectrum is random). A "better" practice, recommended by the Proteomics Publication Guidelines, is to search a decoy database, to estimate the probability that a match to a random peptide has the same value as the one obtained for the identified parent peptide. Basically, the idea is to learn the null-hypothesis distribution of the scores of random matches from a "decoy" database of wrong peptides, generated by the inversion of the sequence, or involving a more general reshuffling of the residues. Obviously, the conclusions on the reliability of the score, drawn in this way, strongly depends on how accurate the decoy database is.

Other approaches aim at giving a global estimate of the quality of the match, by training a classification algorithm on spectra of known identity [37], to be able to distinguish correct and incorrect matches. Again, this approach depends on the choice of the training database.

As pointed out by Kim and co-workers [39], the need for a decoy database is a consequence of the inability to solve the Spectrum Matching Problem: Given a spectrum S and a threshold T for a scoring function, find the prob-

ability that a random peptide matches S with score greater than T. This quantity, which represents the False Positive Rate, is quite difficult to estimate correctly, and is usually replaced by the False Discovery Rate, which is not a characteristic of the individual spectrum, but rather an average property, i.e. the fraction of incorrect guesses among all identifications with score greater than T.

The situation is perhaps even worse for *de novo* interpretation. Basically, all the score functions to assign a quality to the interpretation are empirically derived, and tested against a test-bed of spectra, whose interpretation is considered reliable. Hence, indications of precision and recall can be given for each method, stating how many times, on average, the method does a good job in guessing the parent peptide that originated the spectra (we will make this statements more precise in Chapter 5). For each spectrum, a list of best scoring sequences is usually produced, along with the corresponding score, but there is no real way to determine how reliable the interpretation is.

Chapter 2

Spectral Interpretation as a Statistical Mechanics Problem

We have seen in the previous chapter the strategy underlying protein sequencing by MS/MS, the interpretation problems that have to be faced, and the present methods to deal with them. We have also seen the pros and cons of the two approaches to MS/MS spectra interpretation: database-search and *de-novo* sequencing, and have discussed the issue of the statistical significance of the score of the reported precursor sequence.

Basically, all sequencing algorithms, both database-searching and *de-novo*, yield as a result the solution with the highest score (or, equivalently, with the lowest energy, if we define an energy as the negative score): from this point of view, the only difference between them is given by the state-space of the search, which is just the database in the former case, or all the sequence space of a given mass in the latter. Therefore, we can say that they find the ground-state, or zero-temperature solution, in the appropriate sequence space. Suboptimal solutions, as proposed in the ranked list, corresponding to the excited states, are not included in the solution, but listed a

part. On the other hand, physical intuition suggests that, given a reasonable energy function to score the match between a proposed sequence and the experimental spectrum, poor quality spectra, with a low signal/noise value (defined in some suitable sense), should correspond to high energy and/or high entropy. Indeed, a high number of excited states at small gap from the ground state, representing alternative solutions with essentially the same energy as the best one, should point at a bad prediction, possibly related to missing peaks, to the presence of many or intense noise peaks, or both.

These observations suggest that the introduction of an artificial temperature in the solutions space, and the study of the resulting thermodynamic equilibrium, could provide some valuable internal indication of the quality of the interpretation, in a sense analogous to the False Predictive Rate mentioned in Section 1.2.3. For instance, despite the fact that, in a finite system, the existence of real phase transition is precluded, we could expect that, for each spectrum, there is a low-temperature phase, where one or a few sequences dominate as candidate solutions, while at high temperature a "paramagnetic" phase exists, corresponding to a mixture of a huge amount of solutions. The temperature of transition between the two regimes, as signalled, for instance, by a peak in the heat capacity, could bring, again, important information on the robustness of the interpretation – the higher the temperature, the more reliable the identification.

In order to exploit the benefits of moving out from the zero-temperature solution, we need first of all a proper way to map the problem of spectra interpretation on that of finding the thermodynamic equilibrium of a suitable physical system, and second, an effective way to actually perform calculations. In the following, we will deal with both issues: first we will introduce a discrete unidimensional system, whose states encode all possible amino-acid sequences of appropriate length. Second, we will define the general form of the energy function of the model in term of just on-site and next-neighbors interaction, involving some constraints on the class of scoring function that can be implemented in the model. We will leave the detailed derivation of a

particular form of the energy function, inspired by Bayesian modelling, for Chapter 4, moving forward to introduce a transfer-matrix technique that allows to calculate exactly the partition function of the model, finding its equilibrium solution. We will finally discuss how the equilibrium results can be mapped back to MS spectra interpretation.

2.1 Encoding the sequence space on a 1D lattice model

A peak in the experimental spectrum gives information on the m/z ratio of the N- or C-terminal ions (if we neglect the possibility of internal fragments) obtained in the fragmentation process during CID: such information is not related to detailed sequence of the fragment ions (even if it can be affected by the presence of some amino-acid, as we will detail below), but just depends on the mass and charge of the whole fragment.

We can use this fact to avoid the book-keeping of a combinatorial number of possible sequences, and to define a physical system whose site variables carry information on the mass and charge, as well as some other overall quantities that we will need to characterize completely the parent peptide.

Despite the fact that the fragment masses are incommensurable, their discretization in terms of a unit mass η, roughly equal to an atomic mass unit, is a reasonable and natural approximation. Therefore, we define a discrete one-dimensional lattice of $M + 1$ sites, numbered from 0 to M, where M is the discretized mass of the parent peptide (more precisely, it is the sum of the masses of the residues that compose it, neglecting the extra groups H and OH at the N- and C- term, and neglecting the extra proton of the precursor MH^+). The lattice spacing η is given by:

$$\eta = \sum_a \frac{m_a^m}{m_a} f(a) = 1.0005022782 \ , \tag{2.1}$$

the weighted mean of the ratio between the mono-isotopic residue mass m_a^m and the discrete mass m_a, where the weight $f(a)$ is the natural abundance of the residue a.

Table 2.1 lists all the residues with their mono-isotopic mass and the corresponding discretized value. Testing the model, in Chapter 5, we will ignore post-translational modification and will deal just with sequences made up of natural amino-acids; however, the following discussion holds valid when PTMs are present, by simply augmenting Table 2.1 with the corresponding modified residues masses and corresponding properties.

Any sequence of total mass M will be identified, in the lattice, by the terminal points of each residue, that will coincide (thanks to the mass discretization) with different lattice sites, situated at a distance corresponding to the residue's mass. For instance, referring to Figure 2.1, we will have that the first residue, A, of mass m_A, starts in 0 and ends at ν_A, while the second residue B, of mass m_B, ends at $\nu_{AB} = m_A + m_B$, etc.

We map the possible amino-acid sequences to model configurations by introducing, for each site, a variable $r \in [0, r_{max}]$, where r_{max} is the biggest residue mass (if post-translational modifications are included, it will correspond to the heaviest species, be it a wild-type or a modified residue). We will implement the following rule (by means of a suitable constraint in the energy function): the only values allowed at ν are $r_\nu = r_{\nu-1} + 1$ or $r_\nu = 0$, the latter just holding when $r_{\nu-1} = m(a) - 1$, for some amino-acid a (natural or modified) from the list of chemical species allowed as component of the parent peptide. It is easy to convince oneself that the above rule, together with the boundary conditions $r_0 = r_M = 0$, can generate each and every possible sequence of total mass M, identifying the sites ν where $r_\nu = 0$ as those where a residue ends, see Fig. 2.1. Notice also that the above constraint involves the values of r at neighbouring sites r_ν, $r_{\nu-1}$, and can be dealt with simply as a next-neighbour interaction. Given one possible configuration of the system, the sites ν where $r_\nu = 0$ represent possible fragmentation sites of the corresponding parent peptide. To match them

a	m_a^m	m_a	$f(a)$	q	l_1	l_2	l_3	l_4
G	57.021	57	7.49	0	0	0	0	0
A	71.037	71	5.22	0	0	0	0	0
S	87.032	87	4.53	0	1	0	0	0
P	97.053	97	5.22	0	0	0	0	0
V	99.068	99	1.82	0	0	0	0	0
T	101.048	101	6.26	0	1	0	0	0
C	103.009	103	4.11	0	0	0	0	0
I	113.084	113	7.10	0	0	0	0	0
L	113.084	113	2.23	0	0	0	0	0
N	114.043	114	5.45	0	0	0	0	0
D	115.027	115	9.06	0	0	0	0	0
Q	128.059	128	5.82	0	0	1	0	0
E	129.043	129	2.27	0	0	0	0	0
M	131.040	131	3.91	0	0	0	0	0
H	137.059	137	5.12	1	0	0	1	0
F	147.068	147	7.34	0	0	0	0	0
Y	163.063	163	5.96	0	0	0	0	0
W	186.079	186	1.32	0	0	0	0	0
K	128.095	128	3.25	1	0	1	0	0
R	156.101	156	6.48	1	0	1	1	1

Table 2.1: List of residues with the corresponding mono-isotopic m_a^m and discrete mass m_a as a multiple of η, along with the residue natural frequency $f(a)$ (in percentage). The likeliness of each residue to accept a charge q or to lose a neutral group (water (l_1); ammonia (l_2); accepting water l_3; urea l_4) is reported on the following columns (0 means not allowed; 1 means allowed).

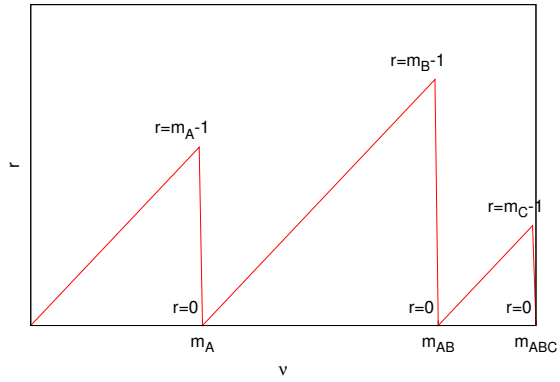


Figure 2.1: Behaviour of the variable r describing the state of the system at the fragmentation site ν, it takes the value 0 at the beginning of the first residue and increases of one at each step of ν. When r reaches the value corresponding to the mass of the first residue A, and then of the second, B, and so on, it is constrained back to the value 0.

against the spectrum peaks, we need to calculate the position m/z of the various ions that can be generated during fragmentation, so that another information that we shall add at each site is the maximal charge that can carry a N-terminal ion (q_ν^N) or a C-terminal one (q_ν^C), according to the parent peptide's charge and the number of residues in the outcoming N- and C-terminal fragments that can loose an electron (K,R,H). The constraints on the charge are a little more complicated than those for the mass variable r_ν (see Appendix A), but again can be written in terms of the variables at sites $(\nu - 1)$ and ν.

The possible fragments generated in a dissociation at a certain position ν also depend on the number of residues that can present neutral losses (such as loss of water, ammonia, phosphate groups, etc.), resulting in a different mass of the corresponding peak. To carry this information, we introduce two vectors of integers $\boldsymbol{l}_\nu^X = \{l_{\nu,1}^X, \ldots, l_{\nu,N_l}^X\}$, $(X = \{N, C\})$, at each site ν, specifying the maximal number of neutral losses of each kind $\boldsymbol{l}_{\nu,\alpha}^N$ that can

be seen in X-terminal fragments. Here $\mathcal{L}$ is the number of neutral losses type considered. The constraints between the values of these vectors at neighbouring site are similar to those for the charge (see Appendix A), with the difference that the maximal number of loss-prone residues in the parent peptide is not known a priori (while the total charge of the precursor peptide is measured by the first Mass Spectrometer, and hence known before the fragmentation process). In the following chapter, we will limit the number of neutral losses considered to two, water and ammonia, for reasons that will be discussed there.

Another important information to carry at each site is one related to the nature of the residue preceding the last one in the peptide sequence: according to the trypsin cleavage rules, a protein chain will not be cut after Lysine (K) or Arginine (R) if the following residue is a Proline (P), or if it is another K or R. So, a $\pi_\nu = 1$ flag will signal that ν lies "inside" a K, R or P residue, so that we must remember this constraint at the closer position $\nu_0 > \nu$ where we will terminate the "current" residue, putting $r_{\nu_0} = 0$. Without this binary flag, sequences containing K and R at arbitrary positions would be considered in the configuration space, enlarging the allowed state space. Actually, this constraint can be softer than the rest, since it is known that sometimes trypsin fails resulting in one (or at most, a few) missing cleavage sites, signalled by some isolated K or R. The above rules are specific to trypsin, which is the most commonly used enzyme in digestion; if a different enzyme is used, different rules should be implemented accordingly.

The last variable that can be used is the number n_ν of residues accumulated in the N-terminal part of the chain with respect to the position ν: this is useful if we want to couple *de-novo* sequencing with database search, or if we are interested in questions as the probability of finding a certain amino-acid as the k-th residue of the parent peptide sequence.

At each ν a fragmentation event can occur and a set of ions can thus be produced; this set, which generally is a small subset of the expected ions, will be matched to the experimental peaks of the target spectrum. An ac-

curate prediction of the fragmentation pattern is not possible, as discussed before, and thus we introduce a binary fragmentation variable $\xi_\nu^{s_i} = 0, 1$ that account for the event of ion formation, taking the value 0 if the corresponding ion species s_i is not produced in the CID, taking the value 1 otherwise.

We collect all the state-variables, characterized above, in a global one, $\sigma_\nu \equiv \{r_\nu, q_\nu^N, q_\nu^C, \boldsymbol{l}_\nu^N, \boldsymbol{l}_\nu^C, \pi_\nu, n_\nu, \boldsymbol{\xi}_\nu\}$, that will contain all the relevant information needed at site ν to yield the correct fragmentation. As detailed in the Appendix A, the constraints imposed on the variables can be resumed in the fact that, if $r_\nu > 0$, σ_ν retains the same values of the state variables at position $(\nu - 1)$ (except $r_\nu = r_{\nu-1} + 1$): the system *remembers* the state it had at the previous site. If $r_\nu = 0$, a new residue is "started" at ν, and the state variables on charges, neutral losses, etc. are updated according to the identity of the residue terminating at ν (that can be recognized by its mass $(r_{\nu-1} + 1)$ and/or state-numbers, with the exception of the couple Isoleucine-Leucine (I-L), see Table 2.1).

The set of all σ_ν, together with their constraints, describe all the possible parent peptides, and specify the relevant information to produce the corresponding fragmentation and ionization patterns. To select which peptide sequence and fragmentation pattern best fits the experimental spectrum, we need an appropriate Hamiltonian, to play the role of the scoring function.

2.2 The Hamiltonian

We will derive and detail an empirical form of the energy function in Chapter 4; here we discuss the general form that such function must have, according to our goals of calculating exactly the corresponding partition function.

The Hamiltonian will consist of two different parts: a first part H^1 represents the energy cost of fragmenting in ν (setting $r_\nu = 0$), and depends only on the experimental spectrum Σ and fragmentation pattern at site ν:

which and how many ions of the different species, together with their neutral losses, are generated and match some peaks. This is a completely local term, acting like an "external field" rewarding or penalizing fragmentation at ν. Due to the local nature, it rules out the possibility of describing non-local events, such as correlations between ion types produced at fragmentations at neighbouring residues, correlations between intensities of fragments from different sites, or the possibility that a peak is actually produced by two different fragmentations, at different sites, that happen to produce ions with the same m/z ratio.

The second part H^2 represents all the constraints, previously mentioned and described in Appendix A, related to the internal consistency of the description of the parent peptide.

H^1 and H^2 can be written as:

$$H^1 = \sum_{\nu=1}^{M} H^1_\nu(\sigma_\nu, \Sigma) \tag{2.2}$$

$$H^2 = \sum_{\nu=1}^{M-1} H^2_{\nu,\nu+1}(\sigma_{\nu-1}, \sigma_\nu) \tag{2.3}$$

The spectrum part of the Hamiltonian $H^1_\nu(\sigma_\nu, \Sigma)$ is the sum of the contributions of the expected fragment ions s_i:

$$H^1_\nu(\sigma_\nu, \Sigma) = \sum_{s_i} H(\nu, s_i, \xi^{s_i}_\nu) \tag{2.4}$$

Here $H(\nu, s_i, \xi^{s_i}_\nu)$ is the energy of matching the theoretical peak of type s_i, generated at the fragmentation site ν, in the target spectrum and is of the form $H(\nu, s_i, \xi^{s_i}_\nu) = \mu + \delta_{\xi^{s_i}_\mu, 1} H^{s_i}_\nu$, where μ is a model parameter representing a chemical potential (see Sec. 4.3 for an accurate characterization of this term).

The interaction part of the energy $H^2_{\nu-1,\nu}(\sigma_{\nu-1}, \sigma_\nu)$ can take two possible values: zero if the corresponding condition is satisfied, and infinity

otherwise:

$$H^2_{\nu-1,\nu}\begin{cases} 0 & \begin{cases} r_\nu \neq 0 \\ r_\nu = r_{\nu-1}+1 \\ q^X_\nu = q^X_{\nu-1} \\ \boldsymbol{l}^X_\nu = \boldsymbol{l}^X_{\nu-1} \\ n_\nu = n_{\nu-1} \\ \pi_\nu = \pi_{\nu-1} \end{cases} \quad \forall X = \{N, C\} \\ 0 & \begin{cases} r_\nu = 0 \\ r_{\nu-1} = m_a - 1 \\ q^X_\nu = q^X_{\nu-1} + \delta^X_q(a) \\ \boldsymbol{l}^X_\nu = \boldsymbol{l}^X_{\nu-1} + \boldsymbol{\delta}^X_l(a) \\ n_\nu = n_{\nu-1} + 1 \\ \pi_\nu = \pi_{\nu-1} + \delta_\pi(a) \end{cases} \quad \forall X = \{N, C\} \\ \infty & \text{otherwise} \end{cases} \tag{2.5}$$

where $\delta^X_q(a), \boldsymbol{\delta}^X_l(a)$ and $\delta_\pi(a)$ take the values -1,0,1 and represent the constraints applied to the first-neighbours interactions. Fragmentation sites are characterized by $r_\nu = 0$ and, in those cases, the values of contiguous variables in $\nu - 1$ and ν have to fulfil the constraints represented by $\delta^X_q(a), \boldsymbol{\delta}^X_l(a)$ and $\delta_\pi(a)$. The latter depend on the amino acid a ending in ν and on the state of the dynamic variables in $\nu - 1$ and ν, as described in the Appendix A.

2.3 Thermodynamics of the model

Using the previous definition of the dynamical variable σ_ν, a state of the unidimensional system is defined as $\phi = \{\sigma_\nu, \nu = 1, \ldots, M\}$. The probability

of the system to visit the state ϕ takes the following form:

$$p(\phi|\Sigma, T) = \frac{e^{-\beta H(\phi,\Sigma)}}{\sum_{\phi'} e^{-\beta H(\phi',\Sigma)}} \tag{2.6}$$

where $H(\phi, \Sigma)$ is the empirical energy function that we use to score the sequences. The quantity at the denominator is the partition function, and is the fundamental quantity to be calculated in equilibrium thermodynamics, since all the others (free-energy, average energy, entropy, etc.) can be related to it. Unfortunately, its evaluation involves the sum over all possible states that the system can assume, which are even more than the possible sequences; considering that for a little peptide of only 800 Da there are 70.440.173 different sequences satisfying the precursor mass constraint, it is easy to understand that a brute-force calculation is unfeasible. We will see in the following section how the calculation can be performed in an efficient way.

Once defined the Hamiltonian one can, in principle, compute interesting values in the form of statistical expected values:

$$U = \langle H \rangle = -\frac{\partial}{\partial \beta} \ln Z = \sum_{\phi} H(\phi, \Sigma) e^{-\beta H(\phi,\Sigma)} \tag{2.7}$$

$$S = \beta \left(\langle H \rangle - F \right) = -\sum_{\phi} p(\phi) \ln p(\phi) \tag{2.8}$$

$$C_V = \frac{1}{T^2} \left(\langle H^2 \rangle - \langle H \rangle^2 \right) \equiv \frac{\partial \langle H \rangle}{\partial T} = T \frac{\partial S}{\partial T} \tag{2.9}$$

Finally, we can derive the probability of a sequence $\widetilde{P}$ through the calculation of the marginals:

$$p_\nu(a) = \langle \delta_{r_\nu,0} \; \delta_{r_{\nu-1},m(a)-1} \rangle \tag{2.10}$$

which represents the probability of the residue species a to end in ν ($r_\nu = 0$), while at $\nu - 1$ the r counter has reached a value corresponding to the discretized mass of the residue a.

We will come back to the explicit calculation of above quantities in the following sections.

2.3.1 The Transfer-Matrix Method

With the definition of H^1 and H^2, the system presents only first-neighbour interactions as only H^2 presents a correlation between adjacent subsystems in $\nu-1$ and ν. In this situation it is possible to resort to the transfer-matrix method, in which the partition function $\mathcal{Z}$ can be calculated incrementally. We write $\mathcal{Z}$ as:

$$\mathcal{Z} = \sum_{\substack{\{\sigma_\nu\} \\ \nu=1\ldots M}} \exp(-\beta \sum_{\nu=1}^{M} H^1(\sigma_\nu) - \beta \sum_{\nu=1}^{M} H^2(\sigma_{\nu-1}, \sigma_\nu))) \tag{2.11}$$

$$= \sum_{\{\sigma_\nu\}} \prod_{\nu=1}^{M} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)} \tag{2.12}$$

where $H_{\nu-1,\nu}(\sigma_{\nu-1}, \sigma_\nu) = H^1_\nu(\sigma_\nu) + H^2_{\nu-1,\nu}(\sigma_{\nu-1}, \sigma_\nu)$, and the first term on the N-terminal edge is taken as $H^1(\sigma_0) = 0$.

We introduce, then, the reduced partition function $\mathcal{Z}_\mu$ that refer to the partition function of the subsystem limited to values of ν in $[0, \mu]$ as in the following:

$$\mathcal{Z}_\mu = \sum_{\substack{\{\sigma_\nu\} \\ \nu=1\ldots\mu}} e^{-\beta H} = \sum_{\substack{\{\sigma_\nu\} \\ \nu=1\ldots\mu}} \prod_{\nu=1}^{\mu} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)} \tag{2.13}$$

$$= \sum_{\sigma_\mu} W_\mu(\sigma_\mu) \tag{2.14}$$

where we have introduced the vector $W_\mu(\sigma_\mu)$ defined as:

$$W_\mu(\sigma_\mu) = \sum_{\substack{\{\sigma_\nu\} \\ \nu=1\ldots\mu-1}} \prod_{\nu=1}^{\mu} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)} \tag{2.15}$$

We can actually write the vector $W_\mu(\sigma_\mu)$ as a function of the previous values at $\mu-1$, $W_{\mu-1}(\sigma_{\mu-1})$:

$$\begin{aligned} W_\mu(\sigma_\mu) &= \sum_{\sigma_{\mu-1}} \left(\sum_{\substack{\{\sigma_\nu\} \\ \nu=1\ldots\mu-2}} \prod_{\nu=1}^{\mu-1} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)} \right) e^{-\beta H_{\mu-1,\mu}(\sigma_{\mu-1},\sigma_\mu)} \\ &= \sum_{\sigma_{\mu-1}} W_{\mu-1}(\sigma_{\mu-1}) e^{-\beta H_{\mu-1,\mu}(\sigma_{\mu-1},\sigma_\mu)} \end{aligned} \tag{2.16}$$

The above expression can be seen as a vector equation, if we introduce a state vector at siteν, $\mathbf{W}_\nu$, whose components are labelled by the possible states of the variable σ_ν, and a Transfer Matrix $\mathbf{T}_{\nu,\nu-1}$ between sites $(\nu-1)$ and ν, of elements $(T_{\nu,\nu-1})_{\sigma_\nu,\sigma_{\nu-1}}$, that carries the information between neighbouring sites; namely.

$$\mathbf{W}_\nu = \mathbf{T}_{\nu,\nu-1}\mathbf{W}_{\nu-1} \ ,$$

which justifies the name of "transfer matrix method" of this section.

Due to the exponential function, at low temperature the above quantities can become very large, so that to avoid computation problems in the implementation of the algorithm as a computer code, we can introduce the normalized $\zeta_\mu(\sigma_\mu)$ as:

$$\zeta_\mu(\sigma_\mu) = \frac{1}{Z_\mu} W_\mu(\sigma_\mu) = \frac{Z_{\mu-1}}{Z_\mu} \sum_{\sigma_{\mu-1}} \zeta_{\mu-1}(\sigma_{\mu-1}) e^{-\beta H_{\mu-1,\mu}} \tag{2.17}$$

We define, then, $\phi_{\mu-1,\mu}$ as the ratio between the reduced partition function of the μ subsystem and the reduced partition function of the $\mu-1$

subsystem. Given the following relations:

$$\begin{aligned} Z_\mu &= \sum_{\sigma_\mu,\sigma_{\mu-1}} \left[W_{\mu-1}(\sigma_{\mu-1}) \left(e^{-\beta H_{\mu-1,\mu}} - \delta_{\sigma_\mu,\sigma_{\mu-1}} \right) + \delta_{\sigma_\mu,\sigma_{\mu-1}} W_{\mu-1}(\sigma_{\mu-1}) \right] \\ &= \sum_{\sigma_\mu,\sigma_{\mu-1}} W_{\mu-1}(\sigma_{\mu-1}) \left(e^{-\beta H_{\mu-1,\mu}} - \delta_{\sigma_\mu,\sigma_{\mu-1}} \right) + Z_{\mu-1} \\ &= Z_{\mu-1} \left[1 + \sum_{\sigma_{\mu-1}} \frac{W_{\mu-1}(\sigma_{\mu-1})}{Z_{\mu-1}} \left(\left(\sum_{\sigma_\mu} e^{-\beta H_{\mu-1,\mu}} \right) - 1 \right) \right] \end{aligned} \tag{2.18}$$

one can write:

$$\phi_{\mu-1,\mu} = \frac{Z_\mu}{Z_{\mu-1}} = 1 + \sum_{\sigma_{\mu-1}} \zeta_{\mu-1}(\sigma_{\mu-1}) \left(\left(\sum_{\sigma_\mu} e^{-\beta H_{\mu-1,\mu}} \right) - 1 \right) \tag{2.19}$$

Following the same method used to calculate the value of the partition function we can calculate the value of other important thermodynamics quantities.

2.3.2 Average Energy

Starting from the definition of $\langle \beta E \rangle = \mathcal{Z}^{-1} \sum_{\sigma_\nu} H \exp(-\beta H)$ that we can rewrite as:

$$\langle \beta E \rangle = \frac{1}{Z} \sum_{\substack{\{\sigma_\nu\} \\ \nu=1\ldots M}} \left(\sum_{\alpha=1}^{M} \beta H_{\alpha-1,\alpha}(\sigma_{\alpha-1}, \sigma_\alpha) \right) \prod_{\nu=1}^{M} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)} \tag{2.20}$$

we notice that we can introduce the mean energy $\langle \beta E_\mu \rangle$ for the subsystem $\nu = 0 \ldots \mu$ as:

$$\langle \beta E_\mu \rangle = \sum_{\sigma_\mu} \varepsilon_\mu(\sigma_\mu) \tag{2.21}$$

where, as in the case of $\zeta_\mu(\sigma_\mu)$, we have that:

$$\varepsilon_\mu(\sigma_\mu) = \frac{1}{Z_\mu} \sum_{\substack{\{\sigma_\nu\} \\ \nu=1\ldots\mu-1}} \left(\sum_{\alpha=1}^{\mu} \beta H_{\alpha-1,\alpha}(\sigma_{\alpha-1}, \sigma_\alpha) \right) \prod_{\nu=1}^{\mu} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)} \tag{2.22}$$

and with a little of maths we can write $\varepsilon_\mu(\sigma_\mu)$ as a function of its predecessor $\varepsilon_{\mu-1}(\sigma_{\mu-1})$

$$\varepsilon_\mu(\sigma_\mu) = \frac{Z_{\mu-1}}{Z_\mu} \sum_{\sigma_{\mu-1}} \Big[\varepsilon_{\mu-1}(\sigma_{\mu-1}) + \\ + \beta H_{\mu-1,\mu}(\sigma_{\mu-1}, \sigma_\mu) \zeta_{\mu-1}(\sigma_{\mu-1}) \Big] e^{-\beta H_{\mu-1,\mu}} \tag{2.23}$$

The final value of $\langle \beta E \rangle$, can then be calculated from the value of $\varepsilon_M(\sigma_M)$ in the following way:

$$\langle \beta E \rangle = \langle \beta E_{\mu=M} \rangle = \sum_{\sigma_M} \varepsilon_M(\sigma_M) \tag{2.24}$$

2.3.3 Heat Capacity

To calculate the heat capacity $C_V = \langle \beta^2 E^2 \rangle - \langle \beta E \rangle^2$, we have to calculate the mean of the square energy:

$$\langle \beta^2 E^2 \rangle = \langle \beta^2 E^2_{\mu=M} \rangle_{(1\ldots M)}$$

We introduce the μ-subsystem mean square energy $\langle \beta^2 E^2_\mu \rangle$ and $c_\mu(\sigma_\mu)$

as before:

$$\langle \beta^2 E_\mu^2 \rangle = \frac{1}{Z_\mu} \sum_{\substack{\{\sigma_\nu\} \\ \nu=1\ldots\mu}} \left(\sum_{\alpha=1}^{\mu} \beta H_{\alpha-1,\alpha}(\sigma_{\alpha-1}, \sigma_\alpha) \right)^2 \cdot \prod_{\nu=1}^{\mu} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)}$$
$$= \sum_{\sigma_\mu} c_\mu(\sigma_\mu) \tag{2.25}$$

where $c_\mu(\sigma_\mu)$ can be calculated recursively:

$$c_\mu(\sigma_\mu) = \frac{Z_{\mu-1}}{Z_\mu} \sum_{\sigma_{\mu-1}} \Big[c_{\mu-1}(\sigma_{\mu-1}) + +2\varepsilon_{\mu-1}\beta H_{\nu-1,\nu} + \\ + \zeta_{\mu-1}(\beta^2 H_{\mu-1,\mu}^2) \Big] e^{-\beta H_{\mu-1,\mu}(\sigma_{\mu-1},\sigma_\mu)} \tag{2.26}$$

The dependence of the C_V on the temperature can show the presence of a transition between different phases or regimes. If the C_V presents a peak at the temperature T_m we can distinguish the passage between a predictive, energetic regime, where the state ϕ is anchored to the spectrum, to a high temperature, entropic regime, where one cannot trust predictions as the system experiments large fluctuations through the configuration space.

2.3.4 Integration of the ξ and state-variables

The state-variables $\xi_\nu^{s_i}$ depend only on the local state at the fragmentation site ν so one can integrate them out and use an effective Hamiltonian in the calculations. In the previous equations, used to compute the thermodynamic quantities, one can pre-calculate the local integration of the weight of the local state $e^{-\beta H_\nu^1}$ and of the energy $H_\nu^1(\sigma_\nu)$, over the $\xi_\nu^{s_i}$. The resulting system is described by the reduced variable $\tilde{\sigma}_\nu = (r_\nu, q_\nu^X, \boldsymbol{l}_\nu^X, \pi_\nu, n_\nu)$, and use a recalculated energy function. Starting from Eq. 2.12, and the

definition of the local energy $H^1_\nu(\sigma_\nu)$, Eq. 2.4, one can write:

$$e^{-\beta H^{\text{eff}}_{\nu-1,\nu}} = \mathcal{Z}^{\xi}_{\nu} = \sum_{\{\boldsymbol{\xi}_\nu\}} e^{-\beta\left(\sum_{s_i}\mu + \delta_{\xi^{s_i}_\nu,1} H^{s_i}_\nu\right)} = \prod_{s_i}\left(1 + e^{-\beta H^{s_i}_\nu}\right)e^{-\beta\mu} \quad (2.27)$$

where μ is the chemical potential, and analogously we can calculate the effective value of the energy that integrate out the $\xi^{s_i}_\nu$ variables, used in ε_ν and c_ν as:

$$\begin{aligned} E^{\text{eff}}_{\nu-1,\nu}(\tilde{\sigma}_{\nu-1}, \tilde{\sigma}_\nu) &= \frac{1}{\mathcal{Z}_\nu}\sum_{\{\boldsymbol{\xi}_\nu\}}\Big(\sum_{s_i}\mu + \delta_{\xi^{s_i}_\nu,1} H^{s_i}_\nu\Big)e^{-\beta\left(\sum_{s_i}\mu + \delta_{\xi^{s_i}_\nu,1} H^{s_i}_\nu\right)} \\ &= \sum_{s_i}\left(\frac{H^{s_i}_\nu e^{-\beta H^{s_i}_\nu}}{e^{-\beta H^{s_i}_\nu} + 1} + \mu\right) \end{aligned} \quad (2.28)$$

Lets then redefine the dynamical variable σ_ν as the reduced variable $\tilde{\sigma}_\nu$ in all the previous equations. Therefore we have to substitute in the computation of the thermodynamic quantities, the Boltzmann weight, with the integrated form of Eq. 2.27, and the energy $H_{\nu-1,\nu}(\sigma_{\nu-1}, \sigma_\nu)$ with the expression of Eq. 2.28. Pre-integration of ξ variables let us to drastically decrease the configuration space and the resulting algorithm will improve notably the running time.

2.4 Peptide Sequencing from the Equilibrium State

To identify the parent peptide sequence, our goal is to calculate, at every site μ, the probability of a residue a to end in μ, $p_\mu(a)$

The system can experiment a fragmentation in μ if σ_μ and $\sigma_{\mu-1}$ fulfil the conditions:

- $r_\mu = 0$;

- $r_{\mu-1} = m_a - 1 \equiv r^*(a)$;
- $n_\mu = n_{\mu-1} + 1$;

In this case the calculation of the probability is straight forward and can be carried out as in the following:

$$p_\mu(a) = \langle \delta_{r_\mu,0}\ \delta_{r_{\mu-1},r^*(a)} \rangle \tag{2.29}$$

To calculate the above quantity, we introduce the symmetrical counterpart of $\zeta_\mu(\sigma_\mu)$, but calculated iterating down from M, as:

$$\begin{aligned} \tilde{\zeta}_\mu(\sigma_\mu) &= \sum_{\substack{\{\sigma_\nu\} \\ \nu=\mu+1\ldots M}} \frac{\prod_{\nu=\mu+1}^{M} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)}}{\tilde{Z}_\mu} \\ &= \sum_{\sigma_{\mu+1}} \frac{\tilde{Z}_{\mu+1}}{\tilde{Z}_\mu} e^{-\beta H_{\mu-1,\mu}(\sigma_{\mu-1},\sigma_\mu)} \tilde{\zeta}_{\mu+1}(\sigma_{\mu+1}) \end{aligned} \tag{2.30}$$

that, if we define $\tilde{Z}_\mu \equiv \frac{Z}{Z_\mu}$, it can be rewritten in the following way:

$$\tilde{\zeta}_\mu(\sigma_\mu) = \sum_{\sigma_{\mu+1}} \sum_{\substack{\{\sigma_\nu\} \\ \nu=\mu+2\ldots M}} \frac{Z_\mu}{Z_{\mu+1}} e^{-\beta H_{\mu-1,\mu}(\sigma_{\mu-1},\sigma_\mu)} \tilde{\zeta}_{\mu+1}(\sigma_{\mu+1}) \tag{2.31}$$

If we write the operator $F^a_{\mu-1,\mu} = \delta_{r_\mu,0}\ \delta_{r_{\mu-1},r^*(a)}$ as the operator that places the end of the residue a in μ, then:

$$p_\mu(a) = \langle F^a_{\mu-1,\mu} \rangle \tag{2.32}$$

$$= \frac{Z_{\mu-1}}{Z_\mu} \sum_{\sigma_\mu} \sum_{\sigma_{\mu-1}} \zeta_{\mu-1}(\sigma_{\mu-1}) F^a_{\mu-1,\mu} e^{-\beta H_{\mu-1,\mu}} \tilde{\zeta}_\mu(\sigma_\mu) \tag{2.33}$$

The collection of $p_m u(a), \mu = 1 \ldots M$, plays a fundamental role in the identification of the parent sequence: at $T = 0$, the only variables different

from zero will be those corresponding to the fragmentation sites and the residues of the lowest energy sequence, thus allowing an easy readout of it. At higher temperature, they will provide a probability profile, that could be useful to perform a database search as a post-processing. In both cases, to evaluate the quality of the prediction we have to define some quantities that compare our prediction to the provided "true" sequence. To this end we resort to Precision, Recall and their combination F-value. Precision is referred to as the fraction of the predicted events that can be accounted as exactly predicted, prec. $= \frac{TP}{PP}$, where TP represents the number of true positives and PP the predicted positives. The fraction of real events predicted by the model is called recall, rec. $= \frac{TP}{RP}$ where RP is the number of real positives, and the harmonic mean of the two measures is called F-value:

$$F\text{-value} = 2\frac{\text{prec.}\cdot\text{rec.}}{\text{prec.}+\text{rec.}} = 2\frac{TP}{RP+PP} \tag{2.34}$$

We define these quantities in two different ways, to assess the goodness of the whole profile or just the best sequence. In the first case we define the value of PP as $PP = \sum_\nu \sum_a p_\nu(a)$ and analogously $TP = \sum_{\nu\in\mathcal{F}(P)} \sum_a p_\nu(a)$ where P is the supposed "true" sequence and $\mathcal{F}(P)$ its set of fragmentation sites. We define RP as the number of residues of the "real" sequence.

The second definition of Precision and Recall focus on the predicted best sequence, and is specially suited to compare the results of the model prediction to other algorithms. We redefine the quantities as follows: if P' is the sequence prediction of the model, then PP' is the length in residues of this sequence and TP' is the number of correct predicted fragmentation sites. RP' take the same form as the previous RP while precision, recall and F'-value are recalculated from those values.

2.5 T-novoMS: implementation issues

The algorithm described in the previous sections has been implemented in a in-house software called "*T-novoMS*" written in C++. Parameters are stored separately in text files and the software accepts as input a spectrum file in the *.dta format by SEQUEST, which is composed by the precursor mass ($m(MH^+)$) and charge (Q) in the first line, followed by peaks centroids and intensities in the following lines. The software run from the command line and accepts as options the filename of the spectrum, a test sequence to compare to the probability profile, the "real" precursor mass to override the original from the spectrum file, and a different parameter file.

The computation of $p_\nu(a) = \langle F^a_{\nu-1,\nu} \rangle$ requires the application of the iterative calculation from both ends $\nu = 0$ and $\nu = M$ in each site ν through the values of $\zeta_\nu(\sigma_\nu)$ and $\tilde{\zeta}_{\nu+1}(\sigma_{\nu+1})$.

This can be accomplished calculating the value of $\zeta_\nu(\sigma_\nu)$ in a first pass, while calculating the thermodynamic values of $F, \langle E \rangle, C_V$ and S. Then one has to *remember* the values of $\zeta_\nu(\sigma_\nu)$ only in the case of $r_\nu = 0$ as required by the fragmentation operator $F^a_{\nu-1,\nu}$.

The procedure followed to determine the most probable sequence at given temperature is the following: one start from the C-terminal of the protein chain and determine the most probable residue to end in that site $a^* = \arg\max_a p_M(a)$. Once selected the last residue one go to fragmentation site $M - m_a$ which is the sites where the last residue starts and look for the most probable residue ending there, the operation is then repeated until the N-terminal is reached. Notice that this operation is not feasible starting from the N-terminal and ending at the C-terminal because the probability profile $p_\nu(a)$, calculated with this algorithm, represents the probability of the residue a to end in ν and these probabilities are normalized at each fragmentation site ν so that only $p_\nu(a)$ at the same ν can be compared.

If the model mis-predict the total mass of the peptide M than a bias is

introduced in Precision and Recall toward peptides with strong N-terminal ions signals. To avoid this problem we calculate both quantities PP and TP starting from both edges, N- and C-terminals, and the mean is taken.

Chapter 3

Characterization of Phenomenological Spectral Distributions

A Tandem Mass Spectrometry product is represented by a list of measured mass-to-charge m/z ratio of the ions that are observed during the measurement. Those values are accompanied by an intensity which is directly proportional to the number of ions observed. Generally the spectrum comes with further information due to the MS_1 spectrum: the precursor ion mass MH^+ and the precursor ion charge Q. Moreover the product ion peaks come mixed to instrument noise, isotopes and other puzzling signals that complicate the task of assigning each peak to a product ion.

The absolute intensity of a spectrum is mainly related to the amount of precursor peptide in the sample, so that it does not provide useful information for identification. However different kind of product peaks present generally different intensities and different ratios with other ion types. Generally b and y ions present higher intensity compared to the others. Noise

peaks are usually considered to present lower intensity.

Knowing the exact distribution of each ion type in the $(\frac{m}{z}, I)$ plane we can infer the probability of a peak to belong to one type of product ion or another, or to the noise.

Thus our strategy, and the objective of this chapter, is to learn the distribution of the different families of peaks from a database of spectra for which a reliable interpretation has been provided. Using this information we will derive in Chapter 4 an energy function for the unidimensional statistical-mechanics model presented in chapter 2 to infer the precursor peptide sequence from the peaks observed in the experimental spectrum.

3.1 Building the learning database

3.1.1 Preliminary choice of the spectra

We collected a learning set of peaks from different spectra. The latter where extracted from a freely available database deposited on the web, and hosting the work of different research groups. The website offering those data is available on the internet at the following address: www.peptideatlas.org from the Seattle Proteome Center.

Those data where collected by different instruments and, in all the cases, they came with the interpretation carried out by the database search algorithm SEQUEST. The databases we have used are the followings, listed with the experimental instrument employed:

- PAe000032 [65] LCQ-DECA ion mass spectrometer (Thermo-Finnigan) and a micro-electrospray source (Brechbuehler);
- PAe000035 [65] LCQ-DECA (Thermo-Finnigan);
- PAe000142 [47] 2D HPLC coupled with the classical LCQ-ESI ion trap

(Thermo-Finnigan);

- PAe000219 [24] μLC-ESI-MS/MS using LCQ-DECA (Thermo-Finnigan);
- PAe000244 [24] μLC-ESI-MS/MS using LCQ-DECA (Thermo-Finnigan);

3.1.2 Filtering for spectra with a reliable interpretation

In MS/MS, the sequence assignation through *database search* algorithms like SEQUEST or Mascot is affected by a significant number of false positives [38], affecting peptide sequencing with incorrect sequences. There are a number of algorithms written to improve peptide scoring and to better distinguish low quality sequence assignations from reliable ones. Some examples are: Protein Prophet [37, 59], SEQUEST-NORM [44], MASPIC [58]. The selection of a reliable data set, from the published experimental data, is then a central point in a database-learning schema.

q	XCorr cut-off	ΔC_n cut-off
1	1.5	0.3
2	3.0	0.3
3	3.5	0.3

Table 3.1: The table reports cut-off values for XCorr and ΔC_n. We filter MS/MS spectrometry databases based on these values to extract a set of reliable spectra. The XCorr cut-off value is correlated to the precursor charge, to avoid the introduction of biases we choose different values for each charge state.

The spectra of each set come with an associated SEQUEST score *XCorr*. This algorithm uses a cross correlation function to measure the quality of the match between the tandem mass spectrum and an amino acid sequence selected from a sequences database. SEQUEST creates a model of the spectrum expected from the precursor peptide represented by the sequence,

relying on a simple interpretation of the CID fragmentation. This "theoretical spectrum" is matched to the target spectrum calculating the cross correlation between them. The latter depends on the quality if the spectrum and on the quality of the match with the theoretical one[18]. On the other hand this score function depends also on the length of the peptide considered as discussed by MacCoss et al. [44], where the authors describe a improved score introducing a normalizing term defined as the cross correlation of the target spectrum with itself, where the latter represents the best possible match.

We want to extract from the downloaded databases a subset of spectra with reliable sequence assignment and, in order to assure the reliability of the sequence assignment, we introduce a cut-off in the XCorr value, discarding low rated data more likely to be incorrect matching. Charge state of the precursor peptide depends on its length as longer ions are more likely to carry an higher quantity of charge, then, in order to correct XCorr bias, we use a different cut-off for each charge.

SEQUEST interpretation comes also with the measure ΔC_n, a parameter that represents how much the first ranked sequence is more probable compared with the second ranked one. We introduce, then, another threshold on the ΔC_n parameter: only spectra whose interpretation has a ΔC_n higher then 0.3 are accepted (in [82, 83] it has been reported as acceptable a value of 0.1).

Table 3.1 reports the selected cut-off values for XCorr and ΔC_n.

3.1.3 Filtering over-represented precursor peptides

The cell proteome can contain thousands of different types of protein, with heterogeneous expression dynamic range. Proteins expression is variable depending on the cell state and type. During the cell life, different proteins can be expressed in different amounts, and the number of copies of the same

molecule can easily span 5 order of magnitude [3]. A single gene can express the same protein with a dynamic range of 4 order of magnitude between the fully repressed and the fully induced state [21].

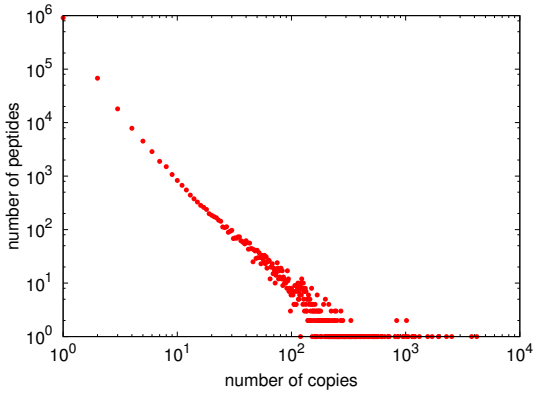


Figure 3.1: Number of different peptides vs number of copies in the downloaded databases. The database presents a large number of copies of the same peptide while the majority of the peptides just come in few copies. Spectra are extracted from the downladed databases on the bases of the XCorr and ΔC_n contraints, before the application of further filters. There are 1.530.738 spectra from 1.021.431 different peptides.

The output databases of MS/MS spectroscopy experiments are composed of spectra of different peptides which present multiple copies of itself, while other peptides just came in very few copies. Figure 3.1 reports the database dynamic range: some peptides are expressed with thousands of copies while the majority just came with few copies.

The presence of few peptides with such number of spectrum copies can invalidate a distribution study. To avoid that the number of copies biases the peaks distribution, we take into account only a maximum number of 10 copies per peptide.

3.1.4 Filtering the Precursor Mass and Sequence Consistency

Despite the available instrument sensibility, it is usual to find erroneous values of the precursor mass, sometimes accountable to human data misinterpretation of the experimental data or parameter. As shown in Fig. 3.2, the difference between the mass provided by the instrument measure of the precursor peak and reported in the spectrum file and the mass expected for the amino acid content of the precursor has a wide distribution covering a range about 6 Da, in the database. This is a source of confusion for the peak interpretation. Notice that the actual distribution presents some peaks uniformly separated by 1 Da that suggests a erroneous selection of the mono-isotopic precursor peak in the experimental procedure, see Subsec. 1.1.4 for a detailed description.

We introduce, then, a further constraint on the experimental precursor mass to be consistent with the theoretical mass calculated from the sequence proposed by SEQUEST. Here the theoretical mass of the sequence is the sum of the mono-isotopic masses of the residues reported with the masses of the N-term, the C-term and the additional hydrogen. The two values of the theoretical and the experimental masses have to differ less then 0.5 Da.

Finally in Table 3.2 we resume the resulting number of spectra that respect all the constraints mentioned above and are considered in this work as the learning database.

3.2 Peak Processing and Recognition

Spectra in the dataset, which came in the SEQUEST data file format, are composed by a list of pairs $\{\pi_\alpha \equiv (\rho_\alpha, I_\alpha)\}$, where ρ_α represents the m/z ratio of the fragment (m is expressed in atomic mass units) and I_α the intensity of its peak, proportional to the ion count. The list of peaks is

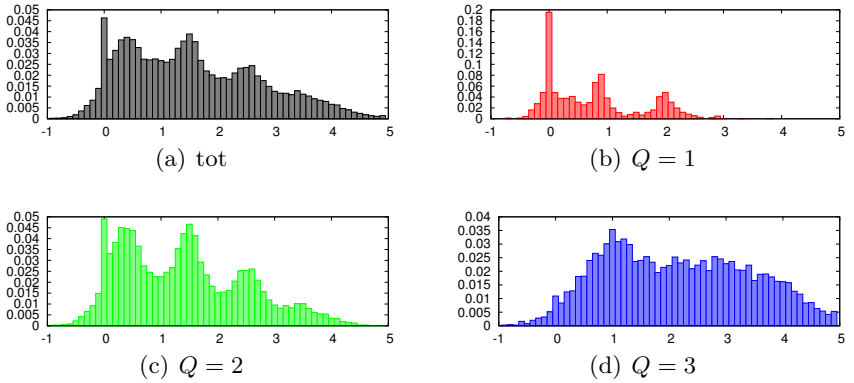


Figure 3.2: Precursor mass distribution as reported in the dta files of experimental data. The reported value is the difference in mass between the value reported in the experimental spectrum file and the monoisotopic value calculated from the sequence proposed by SEQUEST. The figure represent the total distribution (28085 precursor masses from the downloaded spectra, before the application of the precursor mass constraint) (a), and the single distributions accounting for precursors with 1 (b),2 (c) or 3 (c) charges respectively.

q	n. of spectra
1	158
2	7839
3	1390

Table 3.2: Number of spectra for different values of parent charge considered in the learning database. Each precursor ion is present with a maximum spectra number of 10, and respects the mass constrain.

completed with the information about the precursor-ion's mass $m(MH^+)$ and charge Q. The total mass of the precursor ion refers to the mass of the corresponding single charged ion, including the proton carrying the positive charge.

3.2.1 Removing Isotopes

The vast variety of proteins contained in the living organisms is only composed of few types of atoms combined in the amino acid structures. The atomic content of the 20 amino-acid vocabulary is composed by, mainly, carbon (C), nitrogen (N), hydrogen (H) and oxygen (O), with a lower content of sulphur (S).

Those atoms naturally come with different number of neutrons (isotopes), but one can consider that only carbon and nitrogen atoms present a non negligible probability to present as isotope in a precursor ion.

The presence of isotopes in MS/MS mass spectrometry results in multiplicity of peaks corresponding to a single product ion with a typical distribution of peaks at a distance of one Da., or fractions if the ion carry a charge Q higher than one. This set of peaks, called the isotope peak train, is generally observed for the highest peaks and for bigger ions. The first peak represents the isotope-free ion composed only by ^{12}C and ^{14}N, while the

Element	Conc. (Da^{-1})	Isotope	N. A.	Isotope	N. A.
C	0.0436733471	^{13}C	1,07 %		
H	0.0696327313	^{2}H	0.015%		
N	0.0122689921	^{15}N	0,364%		
O	0.0139260967	^{17}O	0.039%	^{18}O	0.201%
S	0.0003500522	^{33}S	0.75 %	^{34}S	4.21%

Table 3.3: The 5 elements representing the basic constituents of the 20 amino acids, with the expected concentration per mass unit in proteins (second column). In nature these elements coexist with a fraction of stable isotopes, reported on the following columns along with their natural abundance (N.A.).

following peaks represent ions with only one or only two isotopes and so on. A train of isotope peaks $\{\pi_\alpha, \alpha = i \dots j\}$ between peaks π_i, π_j, is defined as a set of nearby peaks satisfying the conditions $\rho_\alpha - \rho_i \leq 0.2 \cdot (\alpha - i) Da$, where absent peaks are not allowed. The peak intensity represents the amount of identical ions detected, then the relative intensity of the isotope peak I_α and the mono-isotopic peak I_i depends on the probability to find an isotope-free fragment and the probability to find exactly n isotopes as in the following:

$$\frac{I_i}{I_\alpha} = \frac{p(\mathcal{N}(^{13}C) = 0) p(\mathcal{N}(^{15}N) = 0)}{\sum_{\sigma=0}^{s} \left(p(\mathcal{N}(^{13}C) = \sigma) p(\mathcal{N}(^{15}N) = s - \sigma) \right)} \tag{3.1}$$

where $\mathcal{N}(^{x}X)$ is the number of isotope x of element X in the fragment considered, and:

$$p(\mathcal{N}(^{13}C) = 0) = (1 - p(^{13}C))^{n_C} \tag{3.2}$$

$$p(\mathcal{N}(^{15}N) = 0) = (1 - p(^{15}N))^{n_N} \tag{3.3}$$

$$p(\mathcal{N}(^{13}C) = \sigma) = \binom{n_C}{\sigma} (p(^{13}C))^{\sigma} (1 - p(^{13}C))^{n_C - \sigma} \tag{3.4}$$

$$p(\mathcal{N}(^{15}N) = s - \sigma) = \binom{n_N}{s - \sigma} (p(^{15}N))^{s - \sigma} (1 - p(^{15}N))^{n_N - (s - \sigma)} \tag{3.5}$$

Here $s = (\alpha - i)$ and only isotopes with $s \leq 4$ are considered; $p(^nX)$ is the natural abundance of isotope n of the element X, as in table 3.3, and n_X its the expected number of the element X calculated from the mass-to-charge ratio ρ of the peaks.

When a train of isotopes $\{\pi_\alpha, \alpha = i \dots j\}$ is detected in the spectrum, we keep only the trailing isotope-free peak π_i while peaks containing isotopes are removed.

All the process is repeated for ions carrying a double charge, in the case of precursor ions with charge higher than one.

3.2.2 Normalization and identification of peaks

The theoretical spectrum is composed of peaks calculated from the peptide sequence provided by SEQUEST algorithm. From each peptide bond we calculate all the families f of ions that may have been produced from a collision induced fragmentation at that place. Those six families of ions reflect the fragmentation pattern which present three N-terminal ions $\{a, b, c\}$ and three C-terminal ions $\{x, y, z\}$ (see figure 1.2), whose masses are calculated using the mono-isotopic mass of each amino-acid residue.

For each family f different states of charge q are considered, according to the precursor charge state and sequence composition: in fact Histidine (H), Lysine (K) and Arginine (R) can accept a proton (H^+) and can be found in different charged states. In this work, only ions with charge one or two are considered, that represent the great majority of the outcoming peptides.

Moreover there are some amino acids that can loose a neutral group during CID fragmentation. Usually the loss of water or of an ammonia molecule are the most common; table 3.4 lists all the types of neutral losses l_i we consider. Only ions carrying a number of neutral losses lower or equal to 3 are considered in this work.

i	Type of neutral loss	A.a. involved	Δm
1	water loss (-wat)	S, T	-18.01
2	ammonia loss (-NH_3)	Q,K,R	-17.03
3	water gain (+wat)	H	+18.01
5	urea loss (-urea)	R	-97.98

Table 3.4: Neutral losses types considered during the composition of the theoretical spectrum from the known sequence. The amino acids involved in the loss of those groups and the corresponding mass loss are reported in the following columns.

The theoretical spectrum is then defined by a list of mass-to-charge ratios ρ_i^t that represent all possible ions that the peptide can produce. We don't take in consideration ions related to internal fragmentation, as seen in high energy CID, or Post-Translational Modifications (PTMs). The latter usually act adding a functional group to the protein, which results in residues with modified masses. This can be easily implemented in a *de-novo* algorithm defining the modified residue as a new amino acid and treated separately.

The matching of a peak in the target spectrum $\pi_\alpha = (\rho_\alpha, I_\alpha)$ to the theoretical peak ρ_i is defined on mass proximity basis. Considering that the error on the peak position may be greater at higher values of ρ, angle=-90 we introduce the following definition of matching distance:

$$d(\rho_\alpha, \rho_i^t) \leq \min(\gamma_1, \gamma_2 \rho_\alpha) \tag{3.6}$$

where the parameters are $\gamma_1 = 2.0$, an upper limit to the matching range, and $\gamma_2 = 0.0006$, the relative matching range.

Peak from the spectrum matching a theoretical fragment from the sequence is then tagged with the label representing the corresponding theoretical fragment: $s = (f, q, \vec{l})$.

In each spectrum Σ, peaks span the range $[0 : m(MH^+)] \otimes [0 : I^\Sigma_{\max}]$

which depends on the parent mass and on the expression level of the corresponding peptide, and varies largely between spectra. Since the overall peak intensity is mainly related to the parent peptide abundance, and does not carry relevant information on its identity, we normalize all peaks on the maximum intensity, $\tilde{I}_i = \frac{I_i}{I_{\max}}$. Inspired by the results in [17], we will also normalize on the precursor mass, $\tilde{\rho}_i = \frac{\rho_i}{m(MH^+)}$, reducing all the spectral planes to the common region $[0:1] \otimes [0:1]$.

3.3 Definition of a binning grid

Defining accurate and effective scoring functions is a fundamental step in peptide sequencing techniques, usually accomplished, in the case of *de novo* sequencing, through the analysis of the peaks distribution in reliable databases. Describing fragmentation events and their distribution in the spectrum plane is a hard task usually fulfilled with the definition of a discretized distribution. Many authors have faced the problem in different ways: for instance Frank and Pevzner [23], in their PepNovo algorithm, discretize the normalized space in 20 regions and learn the distribution of intensities of the different fragments (in relation to the intensity of the corresponding y peak). In the HMM algorithm Fischer et al. [22] normalize only the intensities in 5 equi-populated bins.

To find a correct description of the ions distribution in the spectrum we discretize the plane into bins, and build up the histogram of how many peaks fall in each bin. Within a bin, the distribution is considered as uniform. Bin size and number are then modified in order to better reproduce the real peak distribution. The most faithful description of the experimental distribution is, obviously, that with one peak per bin, which, however, represents an over-fitting of the experimental sample, introducing too many parameters, and actually providing little information for modelling. Therefore, it is fundamental to find a model distribution that accurately describe the experimental data with a minimal number of parameters.

To do so, we start by over-fitting the sample data, introducing an huge number of parameters, and successively we reduce them according to a model selection criterion.

The entire plane $[0:1] \otimes [0:1]$ is initially discretized in 12000×12000 regular bins (the integer 12000 has a great number of divisors, which will be useful in the following). In this way, the number of bins exceed the peak population, as shown on table 3.5. To model the overall distribution in the plane, avoiding the over-fitting of the sample, we use the *Bayesian* or *Schwarz Information Criterion* [66](BIC). With this criterion we will select the better discretization of the plane that can describe the sample distribution, maximizing the likelihood estimation between the selected model and the sample, with a limited number of parameters.

As the distribution of the peaks in spectra with different charge differs qualitatively, they are separated in different database and treated separately.

Q	N. peaks	N. matched peaks
1	31666	9071
2	927733	358116
3	173095	80191

Table 3.5: Number of sample peaks in the learning database. We filter the recollected spectra on the quality of their interpretation, and normalize them both in mass-to-charge and in intensity. We report the resulting number of peaks and the number of peaks matching a peptide fragment. Data are reported separately for each considered precursor charge state Q.

The starting model A uses $k(A) = 1.44 \cdot 10^8$ bins to describe the peak distribution. Starting from that, we will introduce a class of other models A_r for the statistical distribution, by dividing $\tilde{\rho}$ axis in r identical intervals, multiples of the basic starting intervals of 12000^{-1}, so that the resulting

number of bins in this case will be $k(A_r) = r \cdot 12000$.

To choose the best model, we use the *Bayesian Information Criterion* as a measure of the quality of the parametric model A_r

$$B(A_r) = -2 \ln \mathcal{L}(A_r) + k(A_r) \ln N_{tot} \tag{3.7}$$

where $\mathcal{L}(A_r)$ is the maximum likelihood, $k(A_r)$ is the number of parameters used to describe the distribution, and N_{tot} the total number of sample data. The maximum likelihood is defined as:

$$\mathcal{L}(A_r) = \prod_{\alpha \in A_r} p((\tilde{\rho}, \tilde{I}) \in \alpha)^{n_\alpha} \tag{3.8}$$

where the variable α represents any bin of the model and n_α its population. A uniform distribution is assumed inside each interval α so that one can write the probability distribution $p((\tilde{\rho}, \tilde{I})|\alpha) = \frac{\delta A}{a_\alpha}$ that depends only on the α-bin area (a_α) and the basic bin area δA ($(1.44 \cdot 10^8)^{-1}$). The probability of a randomly chosen peak to fall into the bin α is then $p(\alpha) = \frac{n_\alpha}{N_{\text{tot}}}$ as the fraction of sample data that fall inside the bin α. One can the write:

$$p((\tilde{\rho}, \tilde{I}) \in \alpha) = p((\tilde{\rho}, \tilde{I})|\alpha) p(\alpha) \tag{3.9}$$

$$= \frac{\delta A}{a_\alpha} \frac{n_\alpha}{N_{\text{tot}}} \tag{3.10}$$

Considering all peaks as independent events, the maximum likelihood $\mathcal{L}(A_r)$ of the model A_r can be computed as:

$$\mathcal{L}(A_r) = \prod_{\alpha \in A_r} \left(\frac{\delta A \; n_\alpha}{a_\alpha \; N_{\text{tot}}} \right)^{n_\alpha} \tag{3.11}$$

$$\mathcal{L}^*(A_r) = \ln(\mathcal{L}(A_r)) = \sum_{\alpha \in A_r} n_\alpha \left(\ln \frac{n_\alpha}{a_\alpha} + \ln \frac{\delta A}{N_{\text{tot}}} \right) \tag{3.12}$$

Our strategy to choose the best discretization will consist in considering all the models A_r; for each of them, we will modify the $\tilde{I}$ discretization

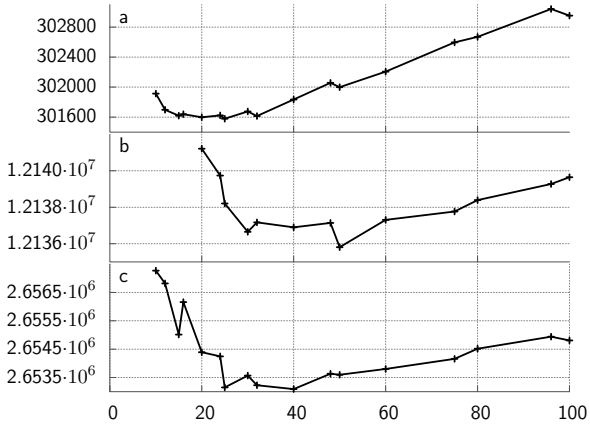


Figure 3.3: *Bayesian Information Criterion* (BIC) dependence on the $\tilde{\rho}$ axis division r. For each value of r, a Simulated Annealing with Montecarlo steps is performed to find the $\tilde{I}$ axis subdivision that minimize BIC. The results are reported for different parent charge states: (a) $Q = 1$, (b) $Q = 2$ and (c) $Q = 3$.

independently for each A_r, finding in each case the binning that minimizes $B(A_r)$. Finally, we will choose the model A^* that minimizes the resulting values of $B(A_r)$ on r. For each A_r, we use a Montecarlo algorithm with Simulated Annealing to change the binning along the intensity axis and minimize $B(A_r)$. We use the following elementary Montecarlo moves

- move a separation edge up or downward;
- remove a separation between two bins, joining two contiguous intervals;
- insert a new division inside a single bin, splitting the interval in two.

The results for the minimal BIC obtained at the end of the simulated an-

nealing procedure for each value of r are reported on figure 3.3. The global optimized subdivision for each precursor charge are the following:

- $Q = 1 \Rightarrow 25$;
- $Q = 2 \Rightarrow 50$;
- $Q = 3 \Rightarrow 40$.

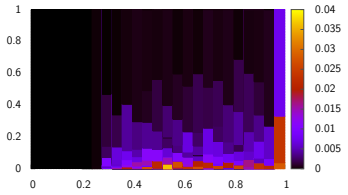
3.4 Selection of Significant Ions

In tandem mass spectroscopy not all the fragments are produced with the same probability: the most frequently observed peaks are mono-charged b and y and, among neutral losses, those of water or ammonia are the most expressed. We want to select between all type of possible product ions a significant subset that is critical to reveal the presence of a fragmentation site.

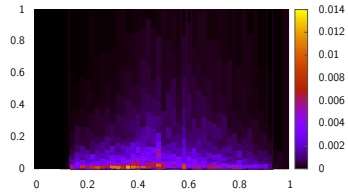
To this end, we analyse the amount of information the presence of a certain type of product ion conveys. We follow a strategy similar to that proposed by Elias et al. [17] in which reduction of Shannon Entropy was used to learn from the database the importance in term of information of each kind of fragment.

Shannon Entropy [68] is a measure of the information content of a random variable or more precisely the uncertainty associated to it. We introduce a model of the system with discrete variables, describing the distribution of the peaks on plane divided in a finite number of bins. The *Shannon Entropy* H of the model $A^*(Q)$ that represents the distribution of peaks from peptides with charge Q, can be defined as

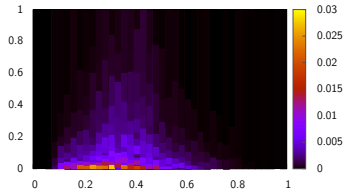
$$H(A^*(Q)) = -\sum_{\alpha \in A^*(Q)} p(\alpha) \log_2 p(\alpha) \tag{3.13}$$



(a) $Q = 1$



(b) $Q = 2$



(c) $Q = 3$

Figure 3.4: Total peaks distribution in the three different parent charge state. The peaks distribution is described in the $[0:1] \otimes [0:1]$ space, subdivided in intervals according to BIC criterion in order to minimize the number of parameters without loosing quality of fit.The results are reported for different parent charge states: (a) $Q = 1$, (b) $Q = 2$ and (c) $Q = 3$.

where the sum is performed over the model bins, and $p(\alpha)$ is the fraction of peaks falling inside the bin α. The separation of the original set of peaks into two subsets increases the information content with the classification of those peaks, the result of such operation can be described with a decrease in the entropy of the entire system. The Shannon Entropy of a general system

separated into A non-overlapping subsets can be written as:

$$H(\alpha|A) = -\sum_{A} p(A) \sum_{\alpha} p(\alpha|A) \log_2 p(\alpha|A) \tag{3.14}$$

The strategy relies on the computation of the entropy decrease due to the recognition/separation of a type of fragment s_i (identified by fragment family, a, b, c or x, y, z, its charge q and its neutral losses $\boldsymbol{l}$) from the rest of peaks. We calculate, then, the Shannon Entropy of the two non-overlapping subsets $\Sigma_{s_i} \equiv \{s_i \in \Sigma\}$, that gather the s_i peaks, and the complementary $\Sigma \setminus \Sigma_{s_i}$. Using Eq. 3.13, this became:

$$H(\Sigma_{s_i}) = \sum_{\alpha} p(\alpha|s_i) \log_2 p(\alpha|s_i) \tag{3.15}$$

$$H(\Sigma \setminus \Sigma_{s_i}) = \sum_{\alpha} p(\alpha|\bar{s}_i) \log_2 p(\alpha|\bar{s}_i) \tag{3.16}$$

where $p(\alpha|s_i)$ and $p(\alpha|\bar{s}_i)$ represent the distributions of the s_i peaks and of all but s_i peaks respectively. The entropy loss $\Delta H(s_i)$ due to the separation/recognition of s_i peaks shows the amount of information gained and can be written as:

$$\Delta H(s_i) = p(\Sigma)H(\Sigma) - p(\Sigma_{s_i})H(\Sigma_{s_i}) - p(\Sigma \setminus \Sigma_{s_i})H(\Sigma \setminus \Sigma_{s_i}) \tag{3.17}$$

where $H(\Sigma)$ is calculated as in Eq. 3.13, $p(X)$ is the fraction of peaks that belong to X.

We consider the fragment type s_i that yield the greatest entropy loss by maximizing $\Delta H(s_i)$ and remove it from the total set of peaks. We then repeat the operation over the remaining peaks. In this way, we obtained the list of product ion types reported in Table 3.6, ranked according to their information content. In the case that a peak matches two fragments, it is assigned to the one with the higher entropy loss.

Fig. 3.5 shows the picture of the entropy loss at every fragment type separation. Notice that the recognition of the firsts few ion types provide

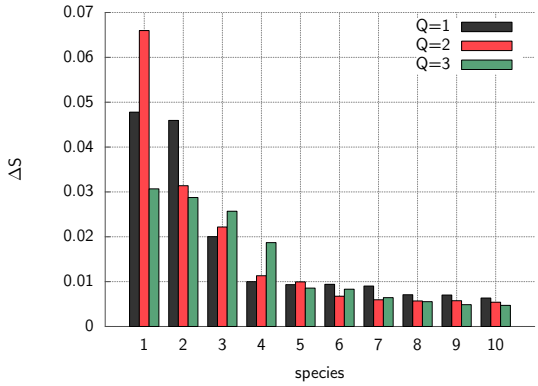


Figure 3.5: Entropy reduction due to interpretation and consequently separation of each species from the rest of peaks. Most of the information is carried by the first few ion species where generally higher positions are occupied by y and b ions, usually with double charged ions or ions with losses of water or ammonia.

the greatest increase in information retrieval, with an high value of entropy loss.

It is remarkable that the usually observed peaks in a general spectrum, and the ions that present higher peaks intensity in experiments, are top ranking on the ordered list of Table 3.6.

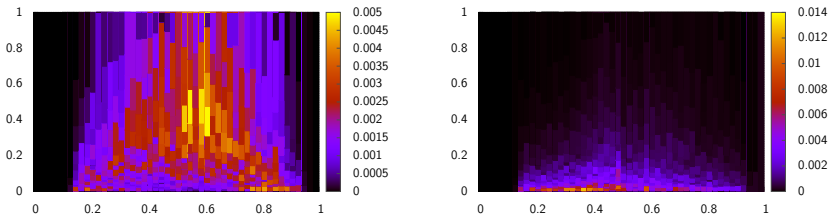
3.4.1 Missing Peaks

Low energy CID fragments the precursor peptide on the peptide bonds but not all the expected ions are expressed on the resulting spectrum. The presence of Proline (P), the three-dimensional structure of the peptide entering the fragmentation chamber as well as other factors, may influence the fragmentation pattern and yield one or several missing peaks.

It is then possible that some expected theoretical peaks, calculated from

seq	$Q=1$		$Q=2$		$Q=3$	
	fragment	N	fragment	N	fragment	N
1	y	1173	y	50208	$y++$	6812
2	b	1003	b	43842	y	6776
3	$b-H_2O$	546	$y++$	14224	$b++$	4723
4	a	515	$y-NH3++$	9825	b	5705
5	$y-H_2O$	369	$b-wat$	21252	$b-H_2O++$	2445
6	$b-2H_2O$	233	$y-H_2O-NH_3++$	6081	$y-NH_3++$	3536
7	$y-NH_3$	556	$y-H_2O$	13385	$b-H_2O-NH_3++$	1044
8	c	321	$x++$	6641	$a++$	2096
9	$b-NH_3$	143	$x-NH_3++$	6546	$y-H_2O-NH_3++$	1798
10	$b+H_2O$	129	$y-H_2O++$	4654	$b-H_2O-NH_3++$	1161

Table 3.6: Ranked list of product ion types, according to the associated *Shannon Entropy* loss. The latter represents the missing information or unpredictability of the random variable, so that the entropy loss can be interpreted as information gain upon their identification. Remarkably, the list reports as top significant those ions that are usually used for spectra identification.



(a) matched (b) unmatched

Figure 3.6: Example of distribution in the normalized and discretized space. The example represents the distribution of peaks corresponding to y-ions (a), and the distribution of the remaining peaks, after the separation (b).

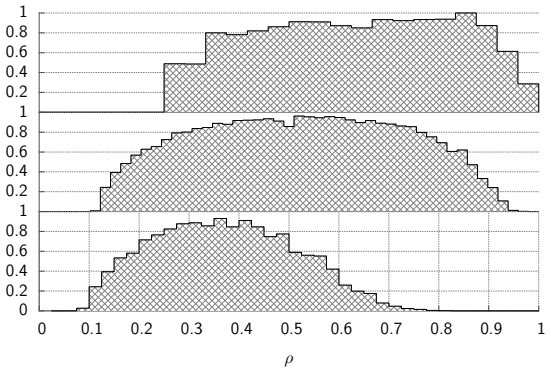


Figure 3.7: Example of the fraction of y ion expressed in those spectra, for peptides with charge $Q = 1$ (top), $Q = 2$ (centre) and $Q = 3$ (bottom), as a function of $\tilde{\rho} \in]0 : 1]$.

the parent peptide sequence associated to any spectrum of the database, do not match any experimental peak in the spectrum, yielding "missing" or "ghost" peaks, i.e., peaks with null intensity.

Figure 3.7 describes the distribution of the expressed fraction of y fragments as function of $\tilde{\rho}$, for doubly charged peptides. The figure shows the instrument limitation at low values of mass-to-charge ratio ρ, while near the peptide centre almost every expected y ion match at least a peak in the corresponding spectrum. The "pit" at the centre of the spectrum with lower expression fraction is due to a post-processing of the output data: in a small window around the precursor ions peaks are removed, if the precursor is double-charged then this window fall in the centre of the spectrum.

Chapter 4

Defining the Potential

In Chapter 2, we have seen that it is possible to map all the possible sequences with the correct total mass on the configurations of a physical model with discrete variables defined on a one-dimensional lattice. We have also seen that, provided the (potential) energy of the system has an appropriate simple form, it is possible to solve the equilibrium of the system and calculate the most likely configuration at low temperature, representing the best parent sequence, and other thermodynamic variables that can be used to estimate the quality of the identification. It is clear that in such approach, the choice of the potential function, to be used as a score for the proposed sequence, is crucial, and different functions can yield very different performances of the overall methods. How should this potential be chosen?

4.1 Definitions and statement of the problem

To answer appropriately to this question, let us start by considering the process of spectra production: the ensemble of identical (and unknown) parent peptides $P*$ undergo Collision Induced Dissociation, which acts as

a statistical process, generating a spectrum of "true" peaks. Meanwhile, a noise source R^{Σ}, that in principle may be different for different samples and different spectra, adds up more peaks, again with an unknown statistical distribution.

Let us call $\Sigma = \{\pi_\alpha, \alpha = 1, \ldots, \mathcal{N}_\Sigma\}$ the total resulting spectrum, as the collection of $\mathcal{N}_\Sigma$ peaks $\pi_\alpha = (\rho_\alpha, I_\alpha)$ with mass-to-charge ratio ρ_α and observed intensity I_α

To estimate the probability that a proposed sequence P is indeed the true precursor ion $P*$, we need to compare the theoretical spectrum associated to P with the experimental one.

The set of theoretical peaks expected in the fragmentation of P depends, in our description, on the location ν_k of the peptide bonds of the precursor peptide in our one-dimensional lattice description: let $\mathcal{F}(P) = \{\nu_k, k = 1, \ldots, L-1\}$, be the set of the possible fragmentation positions of P, with L is the length in residues of P; ν_k is equal to the sum of the discretized masses of the sequence residues up to k.

Let $\mathcal{T}_\nu = \{s_i(\nu), i = 1, \ldots, N_s\}$ be the set of the peaks produced by all kind of fragmentations at the peptide bond located at ν, with N_s the number of expected type of ions per peptide bond, while $\mathcal{T} = \{\mathcal{T}_{\nu_k}, k = 1, \ldots, L-1\}$ is the total set of peaks that can be obtained from P. Notice that N_s depends in general on the fragmentation site, since the number of possible peaks depends on the presence and on the position, in the sequence P, of residues that can be charged or can undergo neutral losses. In our description (see Section 2.2), N_s will depend on the state variable σ_ν: $N_s^\nu(\sigma_\nu)$.

We have seen in table 3.6 in section 3.4 of chapter 3, that CID fragmentation yields several different ion types: a, b, c, x, y, z, with different charges and possible neutral losses. However, it is not expected, a priori, that the full list of ions is produced at every fragmentation site, since the number and intensities of the produced peaks greatly depend on different factors, including the three dimensional structure of the peptide during fragmenta-

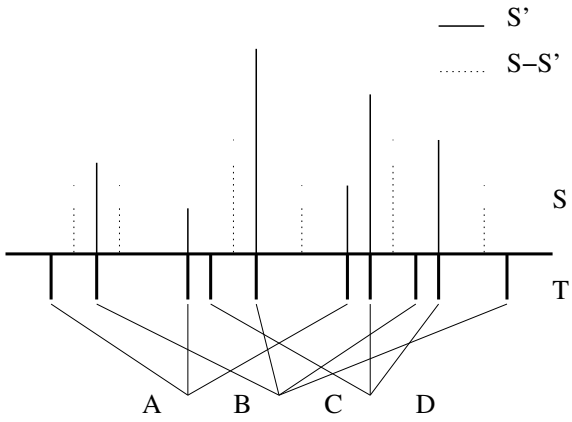


Figure 4.1: Representation of a hypothetical target spectrum Σ of experimental peaks $x_i = (\rho_i, I_i)$, matched by the theoretical peaks of Θ, produced by the amino acid sequence $P = ABCD$. In the figure the subset Σ' represent the experimental peaks matched by at least one theoretical peak.

tion, or the instrument type. For this reason, let's define Θ_ν, a subset of $\mathcal{T}_\nu$, as the ions that have actually been generated from fragmentation at ν, during the experiment. The corresponding collection of ions generated at any position will be $\Theta = \{\Theta_{\nu_k}, k = 1, \ldots, L-1\}$. Any ion in Θ is characterized by a mass-over-charge ratio ρ, representing a theoretical peak position; to test the correspondence between experimental (Σ) and theoretical (Θ) spectra, we define the *image* of a peak $s_i(\nu_k)$ in the spectrum Σ as:

$$\mathcal{I}(s_i(\nu_k)) : \big\{\pi \in \Sigma | d\big(\rho(\pi), \rho(s_i(\nu_k))\big) < d_0\big\} \tag{4.1}$$

where d is the distance between the positions ot the experimental and theoretical peaks (we will simply choose $d(\rho_1, \rho_2) = |\rho_1 - \rho_2\ vert$), and d_0 is a cutoff distance, that we will take to depend linearly on the position of the product ion ρ, since it is a reasonable hypothesis that the errors on a spectral measure are also proportional to the mass. With this assumption

we take $d_0 = min(\gamma_1, \rho(s_i(\nu_k))\gamma_2)$, where the parameters γ_1 and γ_2 are empirically chosen with the values $\gamma_1 = 2.0$ and $\gamma_2 = 0.0006$.

Notice that in Eq. 4.1 the image of a fragment could be represented by several peaks, or also be the null set $\emptyset$. We then define the subset Θ'_ν of Θ_ν, composed of the theoretical ions produced in the fragmentation at ν, that effectively match at least one peak from the real spectrum Σ:

$$\Theta'_\nu = \{s_i(\nu) \in \Theta_\nu | \mathcal{I}(s_i(\nu)) \neq \emptyset\} \tag{4.2}$$

Analogously, the subset of the experimental spectrum Σ of peaks that are the image of the theoretical peaks of Θ'_ν can be defined as:

$$\Sigma'_\nu = \left\{\mathcal{I}(s_i(\nu)) | s_i(\nu) \in \Theta'_\nu\right\} \tag{4.3}$$

The corresponding global sets: $\Theta' = \{\Theta'_{\nu_k}, k = 1, \ldots, L-1\}$ and $\Sigma' = \{\Sigma'_{\nu_k}, k = 1, \ldots, L-1\}$ are defined as before.

We observe that the experimental peaks in Σ' need not necessarily be the real image of the corresponding theoretical fragment: indeed, if several experimental peaks can match the same fragment ion at ν, at most one of them will be the real image of the fragment, the rest being a product of noise. However, there is also the possibility that the ion s_i was not recorded for some reason, and all the experimental matching peaks are a product of noise. Finally, both P and the noise could contribute to the observed intensity of the same peak. To deal with this possibilities, we introduce the concept of *expressed peak*, referring to those ions of Θ'_ν that effectively produced a peak in Σ'_ν:

$$\mathcal{E}(s_i(\nu)) = \begin{cases} 1 & \text{if } \exists\ \pi \in \Sigma'_\nu | s_i(\nu) \to \pi \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

and we define last two subsets of Σ and Θ with the expressed ions and the corresponding experimental peaks: (even if we might not be able to identify

them among the others):

$$\Theta''_\nu = \left\{ s_i(\nu) \in \Theta'_\nu | \mathcal{E}(s_i(\nu)) = 1 \right\} \tag{4.5}$$

$$\Sigma''_\nu = \left\{ \mathcal{I}(s_i(\nu)) | s_i(\nu) \in \Theta''_\nu \right\} \tag{4.6}$$

To avoid to deal with mixed situations as those described above, we can assume that:

Hypothesis 4.1 *If π is a experimental peak of Σ'', we assume that it is entirely due to the corresponding species $s_i \in \Theta''$*

Also, we assume that:

Hypothesis 4.2 *The chances of multiple matches to the same experimental peak π by different fragment ions are negligible.*

The above defined sets imply that all the peaks in $\Sigma \setminus \Sigma''$ are a product of noise, and that all the fragment ions in $\Theta' \setminus \Theta''$ have not produced any peak in the spectrum.

4.2 Protein sequencing as a an inference problem

Let us consider the problem of spectra interpretation as a Bayesian inference problem. We have seen that three different elements are involved in the MS/MS experiment, P, Σ and R^Σ, that can be considered as random variables, extracted from different distributions. Their joint probability $p(P, R^\Sigma, \Sigma)$ can be written as:

$$p(P, R^\Sigma, \Sigma) = p(\Sigma | P, R^\Sigma) p(P, R^\Sigma) \tag{4.7}$$

$$= p(P | \Sigma, R^\Sigma) p(\Sigma, R^\Sigma) \tag{4.8}$$

introducing the conditional probabilities $p(\Sigma|P, R^\Sigma)$ of observing the spectrum Σ given the parent sequence P and the source R^Σ, and $p(P|\Sigma, R^\Sigma)$ of the parent sequence P given the observed spectrum Σ and the noise R^Σ.

We assume that the events generating spectrum noise are completely independent from the precursor ion and the fragmentation process that produce the true peaks:

Hypothesis 4.3 *R^Σ is uncorrelated from the peptide sequence P: $p(P, R^\Sigma) = p(P)p(R^\Sigma)$.*

We also consider the noise distribution to be the same for every spectra, and to be independent from the *a-priori* distribution of the spectra:

Hypothesis 4.4 *R^Σ is independent from the target spectrum Σ: $R^\Sigma \approx R \quad \forall\Sigma$, and $p(R^\Sigma, \Sigma) \approx p(R)p(\Sigma)$.*

Our goal is to estimate the posterior probability $p(P|R, \Sigma)$ that the peptide P is the precursor ion, given the experimental spectrum Σ and noise distribution R. However, we observe that in any *de-novo* framework the actual sequence P cannot be resolved by the experiment, because of the possible mass degeneracies as between Isoleucine and Leucine: for instance, in our modelling scheme, peptides differing by these residues at the same position would correspond to the same configuration of our one-dimensional system, characterized by the same set of fragmentation sites $\mathcal{F}(P)$ and corresponding theoretical ions we are interested in determining $\mathcal{T}(P)$ from the MS/MS data, since $\mathcal{T}(P)$. On the other hand, once given P, the sets $\mathcal{F}(P)$ and $\mathcal{T}(P)$ are completely determined. Therefore we can write, for any random variable Y:

$$p(P|Y) \equiv p(P, \mathcal{T}(P)|Y) = p(P|\mathcal{T}(P), Y)p(\mathcal{T}(P)|Y) \tag{4.9}$$

This is true in particular for

$$p(P|R, \Sigma) = p(P|\mathcal{T}(P))p(\mathcal{T}(P)|R, \Sigma)\,. \tag{4.10}$$

where we have used the fact that $p(P|\mathcal{T}(P), \Sigma, R) = p(P|\mathcal{T}(P))$, since no extra information about P, with respect to the one contained in $\mathcal{T}(P)$, is carried by Σ or R. So, Equation 4.8, using Equations 4.9, 4.10 and hypothesis 4.4 can be rewritten in the following way:

$$p(P, R, \Sigma) = p(\Sigma|P, R)p(P|\mathcal{T}(P))p(\mathcal{T}(P))p(R) \tag{4.11}$$
$$= p(P|\mathcal{T}(P))p(\mathcal{T}(P)|\Sigma, R)p(\Sigma)p(R) \tag{4.12}$$

Thus, from Eq. 4.12 one can finally write:

$$p(\mathcal{T}(P)|R, \Sigma) = p(\Sigma|P, R)\frac{p(\mathcal{T}(P))}{p(\Sigma)} \tag{4.13}$$

which expresses Bayes theorem, and will be our starting point in the following. Interpreting the spectrum will be equivalent to finding the peptide sequence P' that maximize probability 4.13:

$$P' = \underset{P}{\text{argmax}}[p(\mathcal{T}(P)|R, \Sigma)] = \underset{P}{\text{argmax}}[p(\Sigma|P, R)p(\mathcal{T}(P))] \ , \tag{4.14}$$

where we neglect the denominator, which is the same in all the maximization process. Such P' is the best prediction we can give of the true precursor ion P^*.

In the above expression, $p(\mathcal{T}(P))$ is unknown, but some reasonable assumptions can be made on it, basing, for instance, on the amino acid frequencies, or on some other *a-priori* information on the system under study. The key point is, however, how to estimate the probability $p(\Sigma|P, R)$ that, given a sequence P, (represented by its set of theoretical fragments $\mathcal{T}(P)$ encoded in the state variables of our model) and a source of statistical noise R, we will observe the experimental spectrum Σ.

If the physics of the fragmentation process in CID was well understood, along with the distribution of contaminants and other factors responsible

for the noise peaks of a spectrum, we could calculate, for any individual parent peptide molecule, the probability of fragmentation $p^{\texttt{true}}(s_i|P)$ at any site , involving the generation of any ion species s_i. Such probability, together with the corresponding one, $p(\rho|R)$, that the noise R is contributing at position ρ, would allow to calculate the true $p(\Sigma|P,R)$ for any P Identification would then proceed by comparing the experimental spectrum with the theoretical ones associated to all the possible parent peptides, and choosing the peptide whose theoretical spectrum is most similar to the experimental one (using some suitable measure of similarity, as for instance Kullback-Leibler distance). In this approach, the theoretical spectra would be calculated from first principle, with a "true" Hamiltonian function, and the information from the experimental spectrum would be used a posteriori, in the comparison of the average intensities.

The lack of a suitable *ab-initio* Hamiltonian to describe the fragmentation process prompts for a different approach, where information proceeding from the experimental spectrum itself, or by the typical peak distributions in a database of spectra, is used to derive an ad-hoc energy function, so that the experimental spectrum acts as an "external field" biasing the system towards the most-likely parent peptide sequence.

We will use the latter approach here, which is common to all *de-novo* methods. We will also adopt a different view on the MS/MS spectra generation: instead of considering a spectra as the outcome from an ensemble of molecules, and of describing the fundamental process as the fragmentation of one single molecule into two pieces during CID, we will consider the whole spectrum as the overlap between the signal from a noise source and from the fragmentation in multiple sites of a unique parent sequence, both of them generating peaks "as a whole", with a characteristic position and intensity (ρ, I). In other words, due to our ignorance of the true energy function underlying the spectra generation, we move away from the (physically correct) view of a MS spectrum as the overlap between the weak signals of an ensemble of individual products ions, and adopt a more bioinformatics-

oriented view of a spectrum generated by a unique entity, yielding several ion products at each fragmentation site ν, that are rewarded or penalized according to an external empirical potential, based on our results of the previous chapter, on the distribution of probability of the peaks in the (ρ, I) plane. This approach allows us to reward the presence in the spectrum of several ion species representing fragmentations at the same position along the sequence (e.g, detection of y b, etc. ions originating from fragmentations at the same position ν).

To define the energy function, we need to develop Eq. 4.13 in finer detail.

Let us call $\mathcal{A}^{\Sigma}(P)$ an assignment of the sequence P on the spectrum Σ, specifying which theoretical ions Θ were actually produced, which of them Θ'' were detected, and which were the corresponding peaks Σ'' in the spectrum:

$$\mathcal{A}^{\Sigma}(P) = \{\Theta, \Theta'', \Sigma''\} \equiv \{\Theta, \Theta'', \Sigma''; \Theta \setminus \Theta''\} \tag{4.15}$$

The last expression underlines that, given Θ and Θ'', the set of fragmentations that have not produced a peak in the spectrum is also completely determined as $\Theta \setminus \Theta''$.

The probability $p(\Sigma|P, R)$ can be written as the sum over the different ways that spectrum peaks can be assigned to the sequence P:

$$p(\Sigma|P, R) = \sum_{\mathcal{A}^{\Sigma}(P)} p(\Sigma, \mathcal{A}^{\Sigma}(P)|P, R) = \tag{4.16}$$

$$= \sum_{\mathcal{A}^{\Sigma}(P)} p(\Sigma|\mathcal{A}^{\Sigma}(P), P, R) p(\mathcal{A}^{\Sigma}(P)|P, R) \tag{4.17}$$

where the sum over $\mathcal{A}^{\Sigma}(P)$ involves all possible choices of $\Theta \subset \mathcal{T}$, $\Theta'' \subset \Theta'$ and $\Sigma'' \subset \Sigma'$. As we mentioned, this implies that there may be some theoretical ions $s \in \Theta' \setminus \Theta''$ that are considered to have escaped detection, and if any peaks of the spectrum could match them, the latter are considered as noise. Moreover, this also implies that when a single theoretical ion s

can match more than one spectrum peak, the sum also involves all the way the "true" image of s can be chosen in Σ'. To simplify the approach we introduce two hypotheses:

Hypothesis 4.5 *If an ion species $s \in \Theta'$ can match an experimental peak $\pi \in \Sigma$, then the peak should be interpreted as the image of s.*

Hypothesis 4.6 *If an ion species $s \in \Theta'$ matches more than one experimental peak in Σ, then the true image of s will be the peak π that maximize the probability (at fixed Θ and Θ'').*

On the basis of the hypothesis 4.5, we can say that $\Theta'' \equiv \Theta' \subseteq \Theta$. From the hypothesis 4.6 it follows that the peak set Σ'' is completely determined by the definition of Σ, Θ and Θ''. Following previous hypothesis (4.5 and 4.6), the sets Θ'' and Σ'' in the assignment $\mathcal{A}^{\Sigma}(P)$ remains completely determined once defined the set Θ, so that the sum in Eq. 4.17 is simplified: 4.17 change to:

$$p(\Sigma|P,R) = \sum_{\Theta \in \mathcal{T}(P)} p(\Sigma|\mathcal{A}^{\Sigma}(P),P,R)p(\mathcal{A}^{\Sigma}(P)|P,R) \tag{4.18}$$

The probability $p(\Sigma|\mathcal{A}^{\Sigma}(P),P,R)$, given the assignment Θ, Θ'' and Σ'', corresponds to the probability that the remaining unassigned peaks in $\Sigma \setminus \Sigma''$ are due to noise. Under the hypothesis:

Hypothesis 4.7 *Peaks due to noise are independent random events;*

such probability can be factorized as:

$$p(\Sigma|\mathcal{A}^{\Sigma}(P),P,R) = p(R \to \Sigma \setminus \Sigma''|R) = \prod_{\pi \in \Sigma \setminus \Sigma''} p(R \to \pi|R) \tag{4.19}$$

The probability of the assignment $\mathcal{A}^{\Sigma}(P)$ is independent from the noise events and thus $p(\mathcal{A}^{\Sigma}(P)|P,R) \equiv p(\mathcal{A}^{\Sigma}(P)|P)$.

It would be desirable to propose a factorized form also for the latter probability. However, the events that underlie the assignment of the theoretical peaks in Θ'' to the experimental peaks of Σ'' are correlated and cannot be factorized on each ν: for instance the probability of having a peak corresponding to the ion $s_i(\nu_1)$ and that of a peak from ion $s'_i(\nu_2)$ are not independent. However, we can restrict our attention only to the fragmentation sites of P, and not to all ν, and assume that in this case:

Hypothesis 4.8 *Probabilities of different fragmentation sites are independent and can be factorized on $\nu \in \mathcal{F}(P)$.*

Hypothesis 4.9 *Within a given choice of ion species $\Theta_\nu \subseteq \mathcal{T}_\nu(P)$ produced in ν, the probabilities are independents and can be factorized on $s_i(\nu)$.*

With this hypotheses we can rewrite Equation 4.18 as:

$$p(\Sigma|P,R) = \sum_{\Theta \subseteq \mathcal{T}(P)} \left[\prod_{\pi \in \Sigma \setminus \Sigma''} p(R \to \pi|R) \right] \times \\ \times \prod_{\nu \in \mathcal{F}(P)} \left\{ \left[\prod_{s_i(\nu) \in \Theta'_\nu} p^{\mathrm{loc}}(s_i \to \Sigma''(s_i), s_i \in \Theta_\nu, s_i \in \Theta'_\nu|P) \right] \right. \\ \left. \left[\prod_{t_\nu \in \Theta \setminus \Theta'_\nu} p^{\mathrm{loc}}(t_\nu \to \emptyset, t_\nu \in \Theta_\nu, t_\nu \in \Theta \setminus \Theta'_\nu|P) \right] \right\} \quad (4.20)$$

where $p^{\mathrm{loc}}(*)$ indicates that the probability is normalized in a small neighborhood of the theoretical position $\rho(s_i(\nu))$, according to the fact that, once determined the precursor peptide P, the fragmentation sites $\nu \in \mathcal{F}(P)$ and then the expected ions s_i are completely fixed.

We can then write:

$$p(\Sigma|P,R) = \Bigg\{ \prod_{\nu\in\mathcal{F}(P)} \Bigg[\sum_{\Theta_\nu\in\mathcal{T}_\nu(P)} \prod_{s_i(\nu)\in\Theta_\nu} \Big((1-\delta_{\mathcal{I}(s_i),\emptyset}) w^M(s_i,\mathcal{I}(s_i)) + \\ + \delta_{\mathcal{I}(s_i),\emptyset} w^{\bar{M}}(s_i) \Big) \Bigg] \Bigg\} \times p(R\to\Sigma) \quad (4.21)$$

where

$$w^M(s_i,\mathcal{I}(s_i)) = \frac{p^{\mathrm{loc}}(s_i\to\Sigma''(s_i), s_i\in\Theta_\nu, s_i\in\Theta'_\nu|P)}{p(R\to\Sigma''(s_i)|R)} \quad (4.22)$$

$$w^{\overline{M}}(s_i) = p^{\mathrm{loc}}(s_i\to\emptyset, s_i\in\Theta_\nu, s_i\in\Theta_\nu\setminus\Theta'_\nu|P) \quad (4.23)$$

$$p(R\to\Sigma) = \prod_{\pi\in\Sigma} p(R\to\pi|R) \quad (4.24)$$

we introduce the dynamical variable $\xi_\nu^{s_i}$ that indicate if there fragmentation of the peptide in ν produced the ion s_i or not:

$$\xi_\nu^{s_i} = \begin{cases} 1 & \text{if } s_i\in\Theta_\nu \\ 0 & \text{if } s_i\in\mathcal{T}_\nu\setminus\Theta_\nu \end{cases} \quad (4.25)$$

that let us write equation 4.21 as:

$$p(\Sigma|P,R) = p(R\to\Sigma)\times \Bigg\{ \prod_{\nu\in\mathcal{F}(P)} \prod_{s_i\in\mathcal{T}_\nu(P)} \Bigg[\sum_{\xi_\nu^{s_i}=0,1} \Bigg(\delta_{\xi_\nu^{s_i},0} + \\ + \delta_{\xi_\nu s_i,1} \Big((1-\delta_{\mathcal{I}(s_i),\emptyset}) w^M(s_i,\mathcal{I}(s_i)) + \delta_{\mathcal{I}(s_i),\emptyset} w^{\bar{M}}(s_i) \Big) \Bigg) \Bigg] \Bigg\} \quad (4.26)$$

Finally, we go back to Eq. 4.13 and we make another assumption on the *a-priori* probability of the theoretical fragments:

Hypothesis 4.10 *The probability $p(\mathcal{T}(P))$ of the set of theoretical ions from peptide P can be factorized on individual ions, as independent statistical events; the* a-priori *probability of any fragment s_i is independent from its type:*

Hence,

$$p(\mathcal{T}(P)) = \prod_{\nu \in \mathcal{F}(P)} \prod_{s_i \in \mathcal{T}_\nu(P)} p_{\text{ion}} \tag{4.27}$$

where p_{ion} is a constant for all fragmentation sites ν and species s_i.

In the end, our proposal for the posteriori probability of observing P as the parent peptide, given the experimental spectrum Σ and a model R for the noise, is given by Eq. 4.10 together with Eq. 4.14 and:

$$p(\mathcal{T}(P)|\Sigma, R) \propto p(R \to \Sigma) \times \left\{ \prod_{\nu \in \mathcal{F}(P)} \prod_{s_i \in \mathcal{T}_\nu(P)} p_{\text{ion}} \left[\sum_{\xi_\nu^{s_i}=0,1} \left(\delta_{\xi_\nu^{s_i},0} + \right.\right.\right.$$
$$\left.\left.\left. + \delta_{\xi_\nu^{s_i},1} \Big((1 - \delta_{\mathcal{I}(s_i),\emptyset}) w^M(s_i, \mathcal{I}(s_i)) + \delta_{\mathcal{I}(s_i),\emptyset} w^{\bar{M}}(s_i) \Big) \right) \right] \right\} \tag{4.28}$$

4.3 An empirical Energy Function

Equation 4.28 provides us with a recipe to evaluate the fitness of any sequence P to match the spectrum Σ, as a product over all the fragmentation sites $\nu \in \mathcal{F}(P)$ of P, and over all the possible theoretical fragment ions that could in principle be generated by sequence P, of the probability to generate any fragment, times a term accounting for all the possible patterns in which the ions s_i can be actually generated, and, if they are, weighting the probability to have generated a peak in the spectrum (w^M), or to have produced no peak ($w^{\bar{M}}$). The above expression is further multiplied by the probability $p(R \to \Sigma)$ that the whole spectrum is generated by the noise. The latter term is independent from the parent sequence, so that we can

ignore it in the following, and keep the rest of the expression as a guide to introduce an ad-hoc Hamiltonian for our model, by writing:

$$p(\mathcal{T}(P)|\Sigma, R) \propto \left\{ \prod_{\nu \in \mathcal{F}(P)} \left[\prod_{s_i \in \mathcal{T}_\nu(P)} \left(\sum_{\xi_\nu^{s_i}=0,1} e^{-\beta H(\nu, s_i, \xi_\nu^{s_i})} \right) \right] \right\}, \quad (4.29)$$

where β is a fictitious inverse temperature, and we have introduced:

$$H(\nu, s_i, \xi_\nu^{s_i}) = \mu - \delta_{\xi_\nu^{s_i},1} \Big((1 - \delta_{\mathcal{I}(s_i),\emptyset}) \log(w^M(s_i, \mathcal{I}(s_i)) + \\ + \delta_{\mathcal{I}(s_i),\emptyset} \log(w^{\bar{M}}(s_i)) \Big), \quad (4.30)$$

defining the chemical potential as $\mu = -\log p_{\text{ion}}$.

This corresponds to defining the energy function in terms of the probabilities contained in the expressions of w^M and $w^{\bar{M}}$ in Eq. 4.24, that are, however, unknown. We propose an empirical estimate for such quantities resorting to the results of chapter 3, on the phenomenological distributions obtained by the analysis of a large database of spectra with the associated predicted precursor ion.

Namely, in order to define the probability of matching a real peak $p^{\text{loc}}(s_i \to \Sigma''(s_i), s_i \in \Theta_\nu, s_i \in \Theta'_\nu|P)$ and the probability of a theoretical ion to not match any experimental peak $p^{\text{loc}}(s_i \to \emptyset, s_i \in \Theta_\nu, s_i \in \Theta \setminus \Theta'_\nu|P)$, we express them in term of database frequencies. The former represents the probability that, given the fragmentation site $\nu \in \mathcal{F}(P)$, the dissociation event produced the species s_i and this was detected as a peak in the final spectrum. In a similar way, the latter represents the probability that, given $\nu \in \mathcal{F}(P)$, the species s_i was produced, but for some reason it did not yield any peak in the experimental spectrum. We express these probabilities as:

$$p^{\text{loc}}(s_i \to \Sigma''(s_i), s_i \in \Theta_\nu, s_i \in \Theta'_\nu | P) = \frac{N(s_i, x(\Sigma''(s_i)))}{\sum_{t \in \mathcal{T}_\nu(P)} N^T(t, \mathcal{J}_\delta(\rho(t)))} \tag{4.31}$$

$$p^{\text{loc}}(s_i \to \emptyset, s_i \in \Theta_\nu, s_i \in \Theta \setminus \Theta'_\nu | P) = \frac{N(\bar{s}_i, \mathcal{J}_\delta(\rho(s_i)))}{\sum_{t \in \mathcal{T}_\nu(P)} N^T(t, \mathcal{J}_\delta(\rho(t)))} \tag{4.32}$$

$$p(R \to \pi | R) = \frac{N(R, x(\pi))}{N(R)} \tag{4.33}$$

where $x(\pi)$ represents the cell in which the peak π fall, in the cell the distribution of peaks is taken as constant; $\mathcal{J}_\delta(\rho)$ represents the interval $[\rho - \delta, \rho + \delta]$ and considers all peak intensity values. $N(x, y)$ represents the frequency of peaks matched as x inside the interval $y \equiv y_\rho \times y_I$ and $N(\bar{x}, y_\rho)$ the number of absent x type ions in the interval y_ρ, while $N^T(x, y_\rho)$ represent the expected number of peaks x in the y_ρ interval of the ρ axis accounting for both matched and absent peaks: $N^T(x, y_\rho) = N(\bar{x}, y_\rho) + \sum_I N(x, y_\rho \times y_I)$.

Finally, we note that also the value of the *chemical potential* μ defined above is unknown. We observe that its practical role is to penalize the system against matching the experimental spectrum with too many fragment ions, i.e., using sequences with the highest allowed number of residues (and fragmentation sites), polarizing the result towards the lightest residues, ending up in an excess of Glycines. For this reason, we fix it *a posteriori*, optimizing the agreement between the predicted peptide and the "true" parent, as predicted by SEQUEST, for the spectra in the learning database.

Chapter 5

Results and Discussion

5.1 Testing Methods

To test the algorithm described in the previous chapters, we have applied it to a test set kindly provided to us by Fischer et al. [22], and originally selected by Frank and Pevzner [23], of experimental mass spectra. This test set is composed by 280 spectra of double charged peptides up to 1400 Da. coming from tryptic spectrometry experiments, which is accompanied by a reliable sequence assignment by SEQUEST algorithm[23]. It originates from a 18-protein mixture database by Keller et al. [38] and from the open proteomics database by Prince et al. [63].

We compare the outcoming model prediction to the provided sequence assignment that we refer to as "theoretical sequence". To provide a measure of the prediction correctness, we compute Precision (which represents the fraction of correctly predicted fragmentation sites over the total number of proposed fragmentation sites), Recall (which represents correctly predicted sites over the total number of fragmentation sites in the theoretical sequence) and their harmonic mean, F-value, for both the profile and the

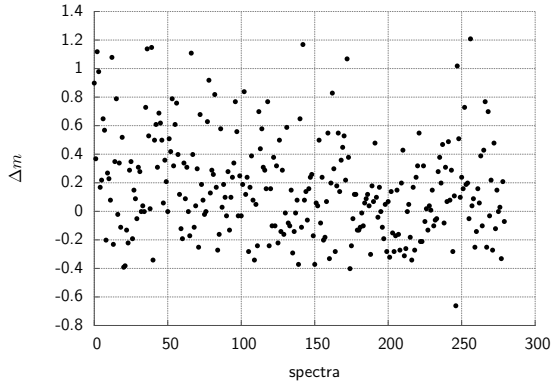


Figure 5.1: Distribution of the difference between the precursor mass reported in the .dta file and the corresponding average mass calculated from the true sequence in the test dataset (Δm). The figure shows that the distribution of the Δm exit the $[-0.5, 0.5]$ range where an exact estimation of the precursor mass can be done by the model.

best sequence as described in Sec. 2.4.

We also probe our method on a prefiltered version of test spectra, removing the peaks that are probably noise, as explained on the following. We also study the effect of enforcing the current parent mass, when the method fails to identify it: as can be seen from Fig. 5.1, there is a mismatch $\Delta m \in [-0.5, 1.2]$ between the mass corresponding to the theoretical sequence and that recovered from the .dta input file, that induced a mass error in 47 out of 280 spectra.

5.2 Results

In the following we present the results of the model implementation and testing: we divide them into results at low temperature, in which the "best

sequence" is extracted and then compared with predictions of other *de-novo* algorithms, and results concerning the temperature dependence of some thermodynamic quantities, that are a feature of our model and can give an insight of the quality of the prediction.

5.2.1 The Low-Temperature Regime: Peptide Identification

The sequencing algorithm resulting from the system described in the previous chapters presents two parameters still to be determined. The first parameter is the number of ion species to use to match the experimental peaks at each fragmentation site, while the second parameter is the value of the chemical potential μ introduced in Sec.4.3. We fix the value of those parameters to the values that optimize the predictions of the algorithm when applied to the learning database.

We selected as working temperature $T = 1$ which was the same temperature adopted when learning the distributions from the database. After some preliminary tests, we saw indeed that this temperature was a good choice, as it is low enough to assure that the system is found basically in the energy minimum, without freezing it in the ground state or provoking numerical problems in the computation.

At this temperature we first tested the software against a selection of spectra from the learning database (all spectra from the database of singly charged precursors, 1000 spectra from the doubly charged and 160 from the triply charged) and determined the number of ions and value of the chemical potential μ that maximizes the F-value separately for each charge value. Here μ can be interpreted as a length limiting parameter as it will discourage configurations of the system that match many peaks just by increasing the sequence length, using smaller residues. Such behaviour would increase the Recall but decrease the value of the Precision. The resulting values for the model parameters are: for singly charged precursors, 3 ions and $\mu(Q = 1) = 0.48$, for doubly charged precursors 3 ions and $\mu(Q = 2) = 0.56$,

and for triply charged precursors 6 ions and $\mu(Q = 3) = 0.28$.

We compared the results of the algorithm with other popular *de novo* sequencing algorithms, such as NovoHMM, Lutefisk, Pepnovo, Pepnovo2011 described in the Subsec. 1.2.2. They were run with the default parameters: NovoHMM was run with the non-grouping option, PepNovo using the CID_IT_TRYP model.

For every algorithm we select the most probable sequence proposed and we compare it with the theoretical sequence provided by SEQUEST. To perform a model comparison we compute TP, PP and RP as well as the F-value for each of the peptides. In Table 5.1 we report the resulting values of precision, recall and F-value. If precision and recall values are very close to each other, then the model is not generating too many fragmentation events in order to match at least a subset (rec.>prec.), neither is producing a low number of events in order not to fail (prec.>rec.). It is fair to notice that Lutefisk and PepNovo algorithms do not predict the entire sequence if they are not sure, so that their recall will be lower than precision, in general.

Table 5.1 shows that the model without modifications does not present high precision or recall values, although, in this particular set, results are better than PepNovo2011, while the others algorithms perform slightly better. It is interesting to notice that the better model in this test is the one that shows a better recognition of the peptide mass. We have introduced some modification in order to identify the model weaknesses. We can force the algorithm to accept an external value of the parent mass (M), to enforce that the mass of the "true" parent sequence is considered, when such mass is not the same as that appearing in the spectrum file. We can also pre-process the spectrum in order to filter out some noise peaks (-pre). The latter filtering is performed by selecting six peaks in each window of 100 Da and discarding the others as noise as in Mo et al. [50]. From the results it is clear that a higher precision in the peptide mass detection will increase noticeably the F'-value, while pre-processing the spectrum to remove the noise peaks does not improve notably the results.

Model	Precision	Recall	F'-value	Mass Mismatch
Lutefisk	0.664	0.717	0.688	43
PepNovo	0.665	0.691	0.676	109
PepNovo 2011	0.589	0.652	0.616	162
HMM	0.778	0.786	0.781	12
T-novoMS	0.657	0.642	0.649	47
T-novoMS-pre	0.663	0.647	0.654	47
T-novoMS-M	0.747	0.732	0.739	0
T-novoMS-pre-M	0.755	0.740	0.747	0

Table 5.1: In this table we report the values of Precision, Recall and F'-value for different algorithms. The last column shows the number of wrongly interpreted precursor masses. True positives are calculated as the mean of true positive accumulated masses from N-term and from C-term, as interpretation with wrong precursor masses can bias the result. The runs of our model have been done with the raw algorithm (*T-novoMS*) or over a pre-filtered spectrum (-pre) or forcing the algorithm to accept the true precursor mass (-M), calculated from the theoretical sequence.

L	*T-novoMS*	*T-novoMS*-pre	*T-novoMS*-M	*T-novoMS*-pre-M	HMM
3	0.743	0.750	0.836	0.836	0.882
4	0.604	0.611	0.704	0.696	0.778
5	0.471	0.500	0.564	0.575	0.664
6	0.368	0.386	0.428	0.439	0.568
7	0.286	0.282	0.335	0.332	0.482
8	0.225	0.196	0.268	0.243	0.410
9	0.139	0.132	0.175	0.171	0.296
10	0.082	0.079	0.114	0.118	0.200

Table 5.2: The fraction of sequence prediction containing an exact string of residues (peptide bond sites) of at least length L.

The same holds true in the application to the learning dataset: there the average F'-value of the identification calculated over a subset of randomly selected spectra for each precursor fragment charge, takes lower values: 0.565, 0.639 and 0.412 for the precursor charge state $Q = 1, 2, 3$ respectively. This value increases if the exact mass of the precursor peptide is passed to the algorithm, giving 0.569, 0.659 and 0.444 respectively.

Tab. 5.2 reports the fraction of predicted sequences that contain a correct subsequence of length greater or equal to L (here we treat the residues only on the basis of their masses and, for example Iso≡Leu). We see that, also in this case, forcing the model to use the original value of the peptide mass improve noticeably the results.

5.2.2 Temperature Dependence and Quality Checks

The dependence on the simulation temperature that governs the overall system behaviour is a specific feature of the presented algorithm and the ability of the model to consider, at each temperature, the weight of all the conformations, that is, of all the possible sequences, is a characteristic absent

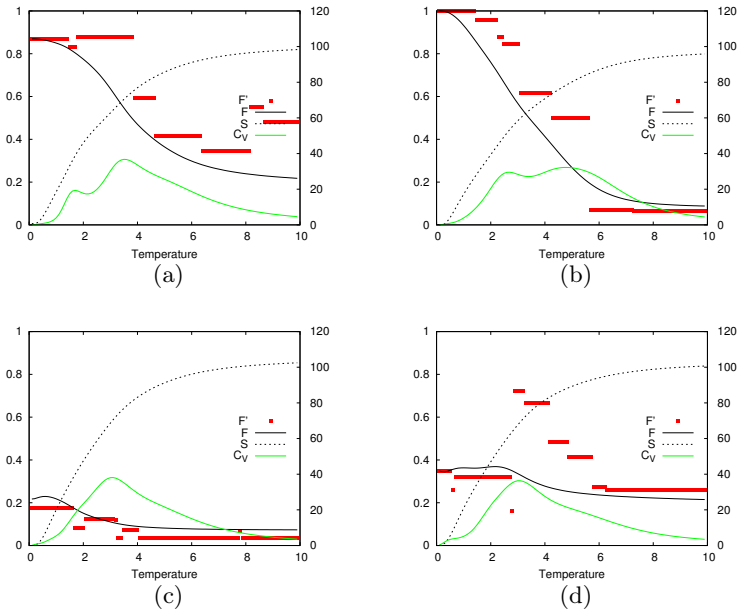


Figure 5.2: Example of results in single spectrum runs. The values of F-value (continuous black line), F'-value (red points), entropy (dotted black line) and specific head (green line) are reported for 4 spectra. The associated "true" sequences are (a) QAIVAEVSEVAK, (b) VVGQLGQVLGPR, (c) EFADNLDSDFK and (d) INALETVTIASK.

in the other *de novo* sequencing algorithms. We will see in the following the effects of the temperature variation on the model predictions.

In a thermodynamic system, the temperature can act as a switch between a energy-driven system behaviour, that reflects the profile of the energy landscape and where the system live in its valleys, and a entropic behaviour where the fluctuations of the system span increasing regions of the configuration space. The analysis of the heat capacity value C_V as a function of the temperature can unveil this behaviour presenting a peak at

a transition temperature where there is a switch between the two regimes.

In Fig. 5.2 the behaviour of some spectra are reported, showing the dependence of F- and F'-values, the heat capacity C_V and the entropy on the system temperature. The same quantities are reported for four different spectra whose "true" amino acids sequences are QAIVAEVSEVAK, VVGQLGQVLGPR, EFADNLDSDFK and INALETVTIASK.

Heat capacity is described in the figure by the green lines. While the first and the second precursors show a three-state-like C_V profile, the third and fourth show only one peak with some smaller deviations. The first behaviour can be explained with the presence of different regions in the precursor sequence that present different stability, which can be interpreted with a bad matching in the spectrum. The peak of the C_V, denoting the transition temperature value, is generally found in the temperature range $[3, 5]$, meaning that for a temperature lower than 3 we can assume that the system is in a energy-driven regime, where the prediction is correlated to the spectrum information, while for a temperature higher than 5 we find the system in a disordered entropy-driven system where fluctuations make impossible every prediction on the most probable sequence.

The behaviour of the F- and F'-values is described by the black continuous line and the red dots respectively, as a function of the temperature. Notice that the first two peptide sequences are predicted with high precision, and indeed at low temperature the values of both F and F' are close to 1, Fig. 5.2(a) and 5.2(b). In the third case the algorithm fails in calculating the precursor mass, and the goodness of match is very low. The fourth peptide just reaches the value of 0.4 at low temperature; the guessed sequence in this case is AATPAEPIPQHK that presents only four fragmentation sites in agreement with those of the theoretical sequence. Notice that the F'-value is a discontinuous function of the temperature, in contrast with F-value which instead is continuous. In fact, while the probability of each sequence change continuously with the temperature, determining the behaviour of F-value, the sequence space is discontinuous

and the passage from a sequence to another change abruptly the number of fragmentation sites matches. Another interesting feature is the fact that in some low-quality predictions, like in Fig. 5.2(d), at higher temperature, near the transition one, the best sequence have a higher F'-value while its F-value is still low. This can be imputable to a bad energy landscape that traps the system into an energy pit, and increasing the temperature, and hence fluctuations, helps the system to escape the energy trap.

The two behaviour regimes are exemplified in the picture of the probabilities profiles at different temperatures. Figure 5.3 shows the probability profile $p_\nu(s_i)$, as described in equation 2.33, in the case of the precursor sequence QAIVAEVSEVAK at four different values of the model temperatures. We stress the fact that the actual probability profile, at non zero temperature, contains the contribution of every sequence in the conformation space compatible with the experimental spectrum. At low temperature, $T = 0.1$, the main contribution comes almost exclusively from the lowest energy sequence state (see Fig. 5.3(a)) and only the fragmentation sites of the most probable sequence show a non-zero probability; moreover the probability of those fragmentation sites is near to 1 with the exception of the first amino acid where K and Q are equally probables as they have the same mass. Increasing the simulation temperature to 1.0, Fig. 5.3(b), the system shows a decrease in stability of small regions of the precursor, in this case the fragmentation sites near N-terminal. At higher temperatures the contribution of high energy configurations becomes important due to the fluctuations of the system and the latter is affected by an higher entropy (see Fig, 5.3(c) and (d)). Notice that some regions, such as the residues VA ranging from 300 to 500 Da and the last pair AK from 1000 Da to the end, of the precursor ion present an higher stability compared to the surrounding regions.

Quality Test. At $T = 0$, the system if found in its ground state, identifying a unique sequence (apart from replacement of residues with identical

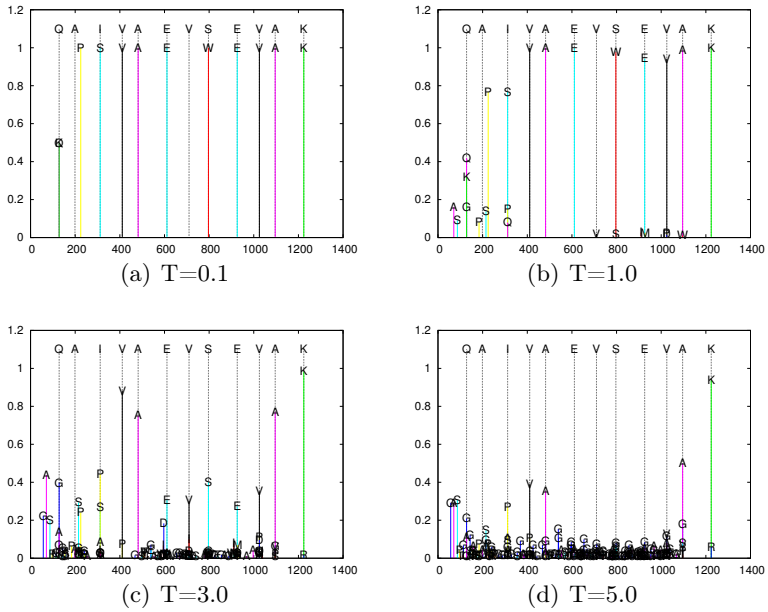


Figure 5.3: Example of a probability profile in the case of low temperature, (a) and (b), and high temperature, (c) and (d). One can notice two regimes energy driven at low temperature and entropy driven at higher temperature. The dotted lines represent the theoretical fragmentation sites calculated from the assigned sequence and are reported as a term of comparison.

mass). So, the algorithm always yields a prediction, and we need to find a way to assess if the latter is reliable or not. Table 5.1 provides information on the average quality of the prediction, but it cannot tell if the identification of one particular spectrum is reliable or not. On the other hand, the possibility to tune the temperature to extract the information, contained in the thermodynamic variables, about the low lying states (that represents identifications with alternative sequences), can provide us with valuable tools to assess the quality of the prediction, a feature that is absent in other *de-novo* and database sequencing algorithms. Therefore, we look for some observable thermodynamic quantities correlating with the F-value, and that can be used as a "proxy" to predict the value of the latter. A possibility could be the position or the height of the peak in the heat capacity: indeed, one could expect that a transition at higher temperatures and/or a high peak, denoting a cooperative behaviour, would reflect a stronger identification at low temperatures, with less competing identifications, and therefore a better identification, provided that the potentials are good enough. Unfortunately, Figure 5.2 suggests that this is not the case, at least with the present potentials: apparently there is no relation between the F-value at low temperatures and the position or the height of the heat-capacity peak, even if the peak temperature can be used to identify the temperature above which the quality of the identification is lost, as the system enters the high temperature, entropy-dominated regime. We have confirmed, by calculating the correlation between peak position or height and F-value at low T, that the former are not good predictors for the latter.

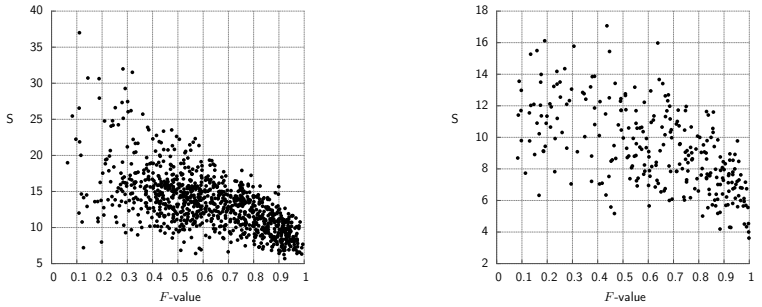
After several attempts, we have found that the quantity that best correlates with the F-value is the entropy at $T = 1$: this is a sufficiently low temperature to allow to identify the best precursor sequence (we have found indeed that it always corresponds to the energy-dominated phase) but has already a reasonable population in alternative conformations to give information about the low-lying structure of the solution space.

Figure 5.4 shows the distribution of the experimental spectra according

to their entropy and F-value, calculated for 1000 randomly selected spectra from the learning dataset, and from the whole test dataset. Several observations can be made: first, the scatter plot reveals that the F-values have a wide range of variability, while it would be desirable that the points were shifted towards the bottom-right corner. These calls for a better definition of the potentials, also including a spectrum-dependent assignment of the chemical potential μ, to tune the length of the predicted peptide to agree with the true one, thus avoiding situations as in Figure 5.3, where at low temperature the peptide length mismatch inevitably lowers the F-value.

A second observation is related to the correlation between the two quantities: accepting that it is not possible to have high F-values for all spectra, the ideal situation would be that of a narrow distribution of the points around a – more or less – linear curve in the F-S plane, so that the knowledge of S would inform quite precisely on the quality of the prediction. The value of the correlation coefficient tells us there is a linear trend in the data, but that the distribution is not very narrow, and Figure 5.3 reveals that while it is relatively easy to find a value of the entropy above which no good interpratation is found, it is difficult to find an upper limit for S, below which the prediction is surely good. As commented in the discussion about Figure 5.2(d), this reflect the existence of some spectra for which the best solution is very stable, very likely and nevertheless wrong, which can be attributed to a limitation is the design of the energy function.

Despite these limitations, important information on the quality of the prediction can be extracted from the data of Figure 5.3. For instance, we can select $F_0 = 0.8$ as a threshold for "good" predictions, and see how the spectra with entropy below (or above) a given threshold are classified according to this criterion. Table 5.3 reports the fraction of the predictions with an entropy below a threshold S_0 or above S_1 that have $F \geq F_0$, along with the number n of spectra in the dataset that fulfil the condition on the entropy. We see, for instance, that if we set $S_0 = 10$ and consider the n =200 spectra (out of a total of 1000) that satisfy the condition $S \leq S_0$,



(a) Learn db (b) Test db

Figure 5.4: Correlation of the system entropy S and the model correctness measure F-value, at temperature $T = 1$. The precursor mass is calculated from the assigned sequence. (a) Data from 1000 random spectra with $Q = 2$ from the learning database (correlation coefficient $r = -0.624$). (b) Data from the test database ($r = -0.598$).

more than 75% of them present an F-value higher than 0.8. In a similar way we can also introduce a second threshold, $S_1 = 14$, above which the population of well-interpreted spectra is very low (less than 3%).

The above results are not yet sufficient to provide the user with a definite knowledge of the value of the prediction, but are indeed a first step towards the definition of an intrinsic quality indicator of the peptide identification, a feature that is missing in other *de novo* sequencing approach. Future development will aim at improving the energy function, so to have a less disperse distribution, and shifted towards higher values of the $F-$value, as well as at characterizing the probability distribution of the F-value versus the entropy, in order to produce reliable confidence intervals to detect false positives and false negatives.

S_0	frac	n	S_1	frac	n
6	1	1	6	0.284	999
7	0.810	21	7	0.274	979
8	0.825	57	8	0.252	943
9	0.818	121	9	0.212	879
10	0.755	200	10	0.168	800
11	0.663	303	11	0.121	697
12	0.598	408	12	0.069	592
13	0.519	507	13	0.045	493
14	0.443	618	14	0.029	382
15	0.391	723	15	0.0072	277
16	0.350	815	16	0	185
17	0.327	872	17	0	128
18	0.314	907	18	0	93
19	0.306	932	19	0	68
20	0.301	946	20	0	54

Table 5.3: Fraction of spectra with $S < S_0$ (second column) or $S > S1$ (fifth column) with F-value above 0.8. The data refer to 1000 random spectra with charge 2, extracted from the learning dataset; column 3 and 6 report how many spectra fulfill the condition specified by the entropy threshold.

Part II

Analysis of the Folding of Myotrophin with the Wako-Saito-Muñoz-Eaton Model

Introduction

In this part of this Thesis we will apply the Wako-Saito-Muñoz-Eaton (WSME) model to the study of Myotrophin, a small ankyrin repeat protein, whose folding equilibrium and kinetics have been recently characterized experimentally.

Modular proteins are frequently found in proteome databases, and they are identified mainly in eukaryotes proteomes, but are abundant also in prokaryotes. The natural abundance of proteins containing duplicated sequence reach 14% [46], which double if we restrict only to eukaryote proteins. In particular the ankyrin motif represents one of the most frequently observed motifs in repeat protein structures, which is composed by two anti-parallel helices followed by a β-hairpin or a long loop (see figure 1). In particular in rats, Myotrophin is found to be related to the heart muscle hypertrophy in which this protein over-express.

Modular proteins drawn researchers interests, providing a different folding paradigm with respect to the well known one of globular proteins, where complex native state geometries, characterized by local and non-local interactions, are most often associated to a simple two-state equilibrium and kinetics. On the contrary, repeat proteins are characterized by tandem arrays of the same structural motif (even if individual repeats show just partial sequence identity, typically, around 25% [40]). Such motifs are usually arranged in a linear fashion, giving rise to elongated structures that may consist of a highly variable number of repeats. Due to this characteristic this kind of proteins are more likely to lack an active site and generally act in the cell processes as building blocks (connective tissue proteins, cytoskeletal proteins...) [46] or as a scaffold for various protein-protein interaction [53].

Interactions in such modular structures take place within a repeat and between adjacent repeats, while truly non-local interactions connecting non-contiguous repeats are lacking. While such organization provide a general-purpose scaffold that can be tuned to bind different species, it is quite

surprising that it is still compatible with a cooperative, two-state folding. Indeed, recent experimental studies have revealed that repeat proteins typically show a two-state equilibrium but a multi-state kinetics [40], driving the attention on the existence of different folding pathways. From a theoretical point of view, repeat proteins provide an ideal framework for modelling and hypothesis-testing, due to their structural modularity, and to the fact that artificial molecules can be built from consensus sequences, so that the role of the different interactions and of the chain length can be dissected and analysed individually.

Not surprisingly, the classical Ising model from statistical mechanics has been used to describe these almost linear systems with local nearest-neighbour interactions, where the spin variables have been identified with individual helices within a repeat [36], or with entire repeats [78], or with the elementary *foldons* identified in a more detailed molecular dynamics simulation [20]. Typically, the external fields and neighbour interaction parameters (h_i and $J_{i,i+1}$ respectively, in their typical textbook denominations) are derived from the experimental analysis of the stability of constructs of different length, and are related to the variation of the areas accessible to the solvent in the folding process.

The identification of the elementary spin variable with a piece of structure as a whole, hinders the possibility to investigate the detailed role of individual contacts between the residues, and of studying the origin of the cooperativity and multi-state kinetics on the residue scale.

Here, we use the Wako-Saito-Muñoz-Eaton (WSME) model [55–57, 75, 76], where the state of each residue i is described by a binary variable $m_i = 0, 1$, representing the unfolded and native state, respectively. Formally, the model differs from the Ising one in that the interactions are not limited to next neighbours, but extend to any distance, provided that the variables corresponding to all the intervening residues are set to the native state. The model equilibrium can be exactly calculated [8, 61, 75, 76], so that energies, free-energies, and fractions of native residues can be easily evalu-

ated. The folding and unfolding kinetics are studied through Monte Carlo simulation, with an elementary step corresponding to the folding/unfolding of one residue.

The model has been applied to describe the folding of many proteins [2, 9, 10, 12, 13, 32, 33, 51, 86], and also to the study of force-induced denaturation of proteins and RNA [11, 28–31]. We apply the WSME model to the study of Myotrophin, a 118 residues protein, ubiquitously expressed in all mammalian tissues [4, 67, 69, 71], made up of four ankyrin repeats. This molecule, a cardiomyogenic hormone, is found to be over-expressed in hypertrophied heart muscle [54]. Its equilibrium has been characterized experimentally as two-state by Peng and co-workers[52] with thermal and chemical denaturation experiments, and later confirmed as such, at least as far as chemical denaturations is concerned, by Lowe and Itzhaki [42], that also studied the kinetics [41, 42]. In the former paper, the authors propose an effective two-state framework to interpret the relaxation kinetics [42] (more precisely, they actually observe some curvature in the unfolding arm of the chevron plot, that can be explained by postulating either a barrier shift or the existence of a high energy intermediate of negligible population).

An extended analysis on several mutants leads them to conclude that, in order to explain within a unique framework the behaviour of both the wild type and the mutants, pathway heterogeneity must be assumed, with the dominant pathway presenting a high energy intermediate, which is lacking in the secondary one [41].

Even if their analysis contains several simplifying assumptions (for instance, the fact that the relaxation rate is just the sum of the rates along the two pathways) they are able to provide very good fits to the experimental data, and to determine that the two pathways present different nucleation sites, on the N-terminal or on the C-terminal part of the protein, respectively. Finally, they show how, by combining mutations, it is possible to make the protein switch between the two pathways.

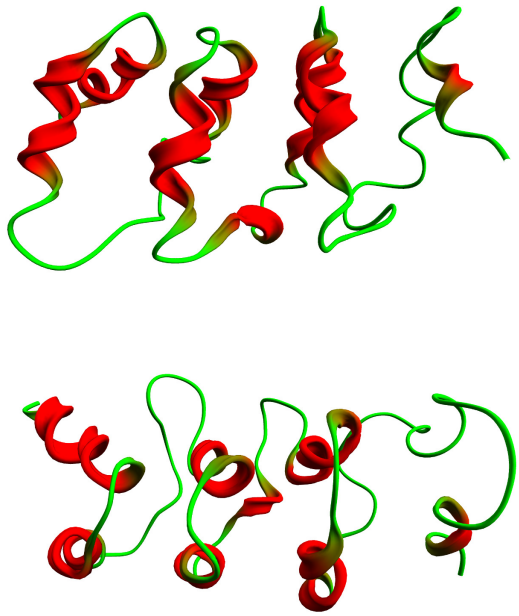


Figure 1: Sketched view of Myotrophin from the side *(top)* and from the top *(bottom)*. The molecule is composed of 118 residue which secondary structure is composed of seven α-helices and one π-helix, ordered in four pairs. Each pair of helices represents an ankyrin repeat. The picture is based on the structure characterized by [81] and published on the Protein Data Bank as 2MYO.

After fitting the model parameters to reproduce the fraction of native protein as a function of the denaturant concentration derived from experiments, we calculate the free energy profiles and relaxation rates, and characterize the relaxation pathways, at low and high denaturant concentrations, for the wild type protein and for a series of mutants, selected to probe different regions and contact distances. We also simulate the set of mutations used in Ref. [41], to test the double pathway hypothesis.

Our goal is to reproduce, at least qualitatively, the experimental behaviour, and to shed light on the nature of the folding nuclei, as well as to recover the role of "pathway switch" played by some mutations. Moreover, we want to clarify the different role played by mutations affecting local or non-local contacts in the same region.

According to the above considerations, the outline of this part is as follows:

WSME model In chapter 6 we review the model introducing a small modification to deal chemical as well thermal denaturation. The system parameters are fitted to agree with the experimental Myotrophin equilibrium denaturation curves in order to test the model against the published data. We introduce and discuss several relevant observables to explain the equilibrium and kinetics results, as well the strategy we use to mimic mutation and to identify the pathway heterogeneity proposed by Lowe and Itzhaki [41]. Thanks to the model characteristics we have a fine control over the folding processes and we can ideally follow each molecule step by step.

Application and Results In chapter 7 we present the results of application of the WSME model to the protein Myotrophin, including comparison with experimental studies for the equilibrium and kinetics data.

We show that the overall two-state-like equilibrium and kinetics of the

wild type protein can hide a non-trivial free-energy profile. These features appear to be related to a careful "design" of the free-energy landscape where the intermediates are always at high energy, so that mutations can alter this picture, stabilizing some intermediates and changing the position of the rate-limiting step.

Furthermore the heterogeneity of folding pathways is qualitatively reproduced as the model predicts two distinct folding pathways, one through the N-terminal and the other through the C-terminal part, even if the variations in the rates upon the experimental mutations cannot be quantitatively reproduced. Interestingly, folding and unfolding pathways appear to be different, even if closely related: a property that is not generally considered in the phenomenological interpretation of the experimental data.

Chapter 6

The Wako-Saito-Muñoz-Eaton Model for Protein Folding

In this chapter we introduce the Wako-Saito-Muñoz-Eaton (WMSE) model and its application to the protein folding problem. After reviewing the main features of the model, we address the calculations of both thermodynamics and kinetics observables that are relevant in experimental studies of the protein folding process, both for ensemble and for single-molecule experiments.

6.1 Methods

6.1.1 Model

WSME is a native-centric model [74], i.e. it relies on the knowledge of the native state of the protein to describe its equilibrium and kinetics. Its binary variables m_k, accounting for the local backbone and side chain angles, describe the state of each residue $k \in [1, N]$ as ordered (native, $m_k = 1$)

and disordered (unfolded, $m_k = 0$). Since the latter state allows a much larger number of microscopic realizations than the former, an entropic cost q_k is given to the ordering of residue k.

The model is described by the effective Hamiltonian (indeed, a free energy, where the solvent and the fast degrees of freedom have been integrated out):

$$H = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\epsilon_{i,j}\Delta_{i,j}\prod_{k=i}^{j}m_k + \sum_{k=1}^{N}(q_kT + \alpha c)m_k, \tag{6.1}$$

where N is the number of residues in the molecule and T the absolute temperature. The product $\prod_{k=i}^{j}m_k$ takes value 1 if and only if all the peptide bonds from i to j are in the native state: indeed, within the model such interaction is ensured only if all the main chain angles of the residues between i and j are in the correct folded conformation. Non–native interactions are disregarded, while native interactions are accounted for in the contact matrix $\Delta_{i,j}$, which counts the number of contacts between atoms of non–contiguous residues i and j in the native structure, according to a cut-off distance criterion. In the following, we will use the contact map calculated from the crystal structure of Myotrophin deposited in the Protein Data Bank (PDB code: 2myo), considering that a contact is established if any two atoms (including hydrogen atoms) from residues i and j are found at a distance less than 3.5 $\mathring{A}$. Figure 6.1 reports the resulting contact map.

The expression above differs from the original one for the last term, accounting for the interaction between the denaturant, usually urea, and the protein backbone (as suggested by Auton et al. [6], and also in agreement with the choice in Ref. [20]), where c represents the urea molar concentration and α is a new parameter.

Setting the Values of the Parameters. For the sake of simplicity, we take homogeneous parameters $\epsilon_{i,j} = \epsilon$, $q_i = q$, for each i and j, with $\epsilon, q > 0$, to model the wild type protein. We use the experimental thermodynamics

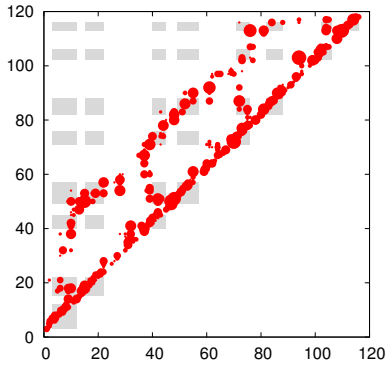
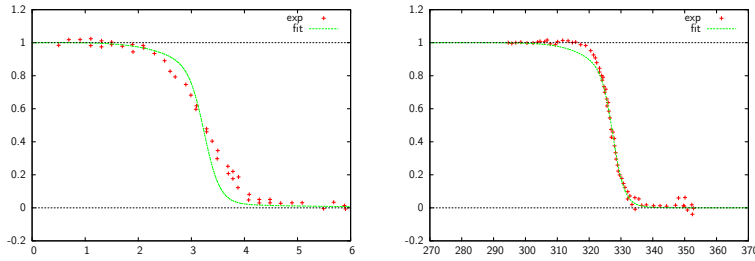


Figure 6.1: Weighted contact map for Myotrophin (2myo). The area of each circle is proportional to the weight of the contact between residues i and j, in terms of the number of inter-atomic interactions. Inter-atomic contacts are established if two atoms belonging to residue i and j respectively are nearer then 3.5 $\mathring{A}$. Darker areas represent contacts between residues that belong to helices.

data to adjust the parameters: we set $q = 2R$, as in [85], and find ϵ, and α by fitting the equilibrium experimental data of Refs. [42, 52] for the wild type species. We first calculate the native and unfolded baselines $n(z)$ and $u(z)$, where z is the temperature or denaturant concentration, respectively, from the data, following the usual experimental procedure, and consider the order parameter:

$$p(z) = \frac{m(z) - u(z)}{n(z) - u(z)}\,, \tag{6.2}$$

normalized between zero (unfolded) and one (native). Here $m(z)$ is the equilibrium average fraction of folded residues, defined in Equation 6.4. Then, we adjust the ϵ parameter, imposing that the temperature T_m at which $p(T_m) = 0.5$ coincides with the experimental mid-folding temperature $T_m = 327$ K . Then, we do the same for the α parameter, imposing that $p(c_m) = 0.5$ at the experimental mid-folding denaturant molar concentration $c_m = 3.2$. The resulting values of the parameters are used in the whole



(a) [Urea] (b) Temperature

Figure 6.2: Fit of the order parameter $p(z)$ from the WSME model to the experimental data for (a) urea-induced denaturation[42] and (b) thermal denaturation[52]. Parameters values are: $\varepsilon = 0.2276$ kJ/mol, $q = 0.0166$ kJ/(K mol), $\alpha = 0.208$ kJ/([urea] mol)

study, for both wild type and mutated species. The results are reported in figure 6.2.

Mutations. Mutations are mimicked by perturbing a group of contacts, of one or more residues as detailed below, through the addition of a $\Delta\epsilon_{i,j}$ to the corresponding interactions. To make comparison easier, the same total perturbation of ε_{tot}=9.21 kJ/mol, comparable with those reported in Ref. [42], is introduced for all mutants, so that the $\Delta\epsilon_{i,j}$ will vary between mutants, according to the number of affected contacts, resulting as follow:

$$\Delta\epsilon_{i,j} = \varepsilon_{\text{tot}} \frac{n^{\text{ct}}_{i,j}}{n^{\text{ct}}_{\text{tot}}} \tag{6.3}$$

where the number of affected contacts between residue i and residue j is $n^{\text{ct}}_{i,j}$ and $n^{\text{ct}}_{\text{tot}}$ is the total amount of affected contacts.

The list of analysed mutations is reported in Table 6.1, and was selected to probe different regions of residues and different distances between contacting residues.

name	description
WT	wild type protein
$S_{1,2}$	contacts of residues [5...10] with residues [17,18]
S_3	contacts of residue 32
$S_{3,4}$	contacts of residues [36...44] with residues [49...56]
$S_{1,4}$	contacts of residues [9...18] with residues [45...53]
$S_{5,6}$	contacts of residues [71...76] with residues [82...88]
$S_{3,6}$	contacts of residues [42...52] with residues [78...83]
S_7	contacts of residue [103] with residues [94...101]
$S_{7,8}$	contacts of residues [104,105] with residues [113,114]
$S_{5,8}$	contact between residue 76 and residue 113

Table 6.1: List of mutations analysed in this work. The same total destabilization of 9.21 kJ/mol was used in all cases, adapting the individual $\Delta\epsilon_{i,j}$ of each contact. The indices m, n in $S_{m,n}$ specify the helices involved in the mutated contacts. S_3 actually affects a loop residue, close to helix 3.

We also analyse the case of some multiple mutations that have been investigated experimentally in Ref. [41] to test the two-pathway interpretation. In order to compare our model to those results, we have simulated the effect of those mutations applying a stabilizing or destabilizing perturbation (whose energy is taken from Ref. [41]) equally spread on all the contacts of the mutated residues. The following mutants are considered:

- E17V/D20L ($\Delta\Delta G_{\text{eq.}} = -7.12 - 2.09$ kJ/mol C: a stabilized mutant)
- A9G ($\Delta\Delta G_{\text{eq.}} = 10.38$ kJ/mol C)
- A115G ($\Delta\Delta G_{\text{eq.}} = 3.39$ kJ/mol C)
- A115G/A9G
- A115G/E17V/D20L
- A9G/E17V/D20L/A115G

Multiple mutations are considered as independent, and the corresponding energetic perturbation is applied to each point mutation separately, so that, e.g., $\Delta\Delta G_{A115G/A9G} = \Delta\Delta G_{A115G} + \Delta\Delta G_{A9G}$.

6.1.2 Thermodynamics

The equilibrium values of all thermodynamic quantities are calculated resorting to the exact solution of the model [8, 61]. In particular, we will study the fraction of native residues:

$$m = \frac{1}{N}\sum_{i=1}^{N}\langle m_i\rangle\,, \tag{6.4}$$

We can introduce the reaction coordinate $\rho = \frac{M}{N}$ as the fraction of native residues M. Free-energy profiles can be defined as a function of ρ:

$$F(\rho) = -RT\log Z_\rho\,, \tag{6.5}$$

where $Z_\rho = \sum'_{\{m_i\}} \exp(-H/RT)$, and the sum $\sum'(\bullet)$ is restricted to the states with a fixed number of native residues $\sum_i m_i = M$. The Z_ρ can be easily calculated within the framework of the exact solution mentioned above.

Finally, we will study the average values

$$\mu_{i,j} = \langle m_{i,j} \rangle \,, \tag{6.6}$$

and

$$\nu_{i,j} = \langle (1 - m_{i-1}) m_{i,j} (1 - m_{j+1}) \rangle \,, \tag{6.7}$$

of the products $m_{i,j} = \prod_{k=i}^{j} m_k$. This product can assume only values one or zero, meaning that the whole string from residue i to residue j is respectively completely native or not. In this framework, $\mu_{i,j}$ represents the equilibrium probability that the region between i and j is native, while $\nu_{i,j}$ represents the probability that the same region is capped by unfolded residues, thus representing an isolated native region.

6.1.3 Kinetics

The kinetic evolution of the model outside the equilibrium is described through a discrete–time master equation, $p_{t+1}(x) = \sum_{x'} W(x' \to x) p_t(x)$, for the probability distribution $p_t(x)$ at time t, where $x = \{m_k, k = 1, \ldots N\}$ denotes the state of the system. Notice that this expression is not amenable to analytical treatment (even if an accurate semi-analytical approximation exists [84, 85]), since by construction $W(x' \to x)$ is a $2^N \times 2^N$ matrix.

In this work the kinetics will be studied by means of Monte Carlo simulations: as in previous works [84, 85], the transition matrix W is specified by a single bond flip Metropolis rule, which implies that a flip is accepted or rejected according to its equilibrium probability, at the temperature and denaturant concentration specified for the simulation. In order to study separately the folding and unfolding kinetics we fix folding condition (T=293.15

K, c=0) and unfolding conditions (T=293.15 K, c=12). The system is prepared in a initial state far from equilibrium and this is defined as a random configuration extracted with the infinite temperature equilibrium probability in the folding case, so that the initial and final fraction of native residues are $m(t = 0) = 0.12$ and $m(t = \infty) = 0.97$ respectively, for the wild type protein (slightly different values are obtained for the mutants). In unfolding simulations, the system is prepared in a fully native state ($m_i = 1$ for each i), while $m(t = \infty) = 0.0047$, for the wild type protein.

We study the relaxation of the average fraction of native residues $m(t)$: at each time, the average is formally calculated as in Equation (6.4), but with the $\langle \bullet \rangle$ now indicating the average over $\mathcal{N}$ single molecule simulations, that is, over an ensemble of $\mathcal{N}$ molecules. We choose $\mathcal{N}$=2000 as a reasonable trade-off between detecting a neat signal and reducing simulation time. We fit $m(t)$ with one- or two-exponential expressions, namely:

$$m(t) = m_{eq}(T, c) + c_1 e^{-k_1 t}\,, \tag{6.8}$$

or

$$m(t) = m_{eq}(T, c) + c_1 e^{-k_1 t} + c_2 e^{-k_2 t}\,, \tag{6.9}$$

where $m_{eq}(T, c)$ is the equilibrium value at the temperature T and denaturant concentration c, obtained from the thermodynamics calculations. The fitting parameters are the rates k_i and the corresponding amplitude c_i.

Pathways Heterogeneity. In order to characterize the folding and unfolding pathways, we rely on the secondary structure formation (denaturation) times. Myotrophin is composed by eight helices (7 α-helices and one π-helix) which are supposed to be partially stable structures, as stabilized by internal hydrogen bonds. These can then be interpreted as fundamental structural motifs. To characterize the helices folding and denaturation in order to find a common behaviour, we define the regions $h_l = (i_l, j_l)$, $l = 1, \ldots, 8$ corresponding to the eight helices of the native Myotrophin structure, as well as the regions $R_{\alpha,\beta}$ encompassing the fragment from helix

α to β inclusive (that is, from residue i_α where helix α begins, to residue j_β where helix β ends). After defining the folding time t_f as the first passage time (in Monte Carlo steps) through the state with all the helices formed ($R_{1,8}$), we identify, for each single molecule simulation of the folding process, the stabilization time $t^{(f)}_{\alpha,\beta}$ of each region $R_{\alpha,\beta}$ as the last time it turns completely native (thus, waiting for all the fluctuations to fade away). This choice is a natural generalization of that proposed in Ref. [86], to the present case with many elements of secondary structure: notice indeed that, due to the model characteristics, the stabilization of $R_{\alpha,\beta}$ in the native conformation is a necessary and sufficient condition for the formation of contacts between helix α and β (if any), as well as between all pairs of helices k,l, with $\alpha \leq k < l \leq \beta$. The determination of $t^{(f)}_{\alpha,\beta}$ for all regions allows us to determine pathways in the secondary structure formation, and to identify two main pathways in the folding and unfolding of Myotrophin (see Chapter 7 below).

We define also the probability of a given region $R_{\alpha,\beta}$ to fold before a different $R_{\alpha',\beta'}$, relying on the folding times as defined before. This can induce interpretation errors in case of wildly fluctuating elements that translate in a measured late stabilization and an artificial ordering of the helices folding times. We have observed, in any case, that the only elements in Myotrophin for which strong fluctuations could induce an interpretation problem are the first and last helix, which on the other hand turn out to be unimportant for pathway determination (see Chapter 7). For the other helices, we observe that local fluctuations can indeed invert the order by which a region is stabilized, in different single-molecule runs, starting from its constituent elements. However, the difference in stabilization times among the latter is small, allowing to group clearly which elements stabilize basically altogether in the folding process.

We do the same for the unfolding simulations: now the unfolding time t_u is defined as the first passage time in a state with $m < 0.09$, and for each single molecule simulation, we record the last time $t^{(u)}_{\alpha,\beta}$ in which each region

$R_{\alpha,\beta}$ switches from the native to unfolded state.

Chapter 7

Application of the WSME model to Myotrophin

In this chapter we report results of the application of the WSME model, described in chapter 6, to the *in silico* characterization of the equilibrium and the kinetics of Myotrophin (pdb code: 2myo), an ankyrin repeat protein exhibiting a three-dimensional structure composed by the repetition of a simple motif, while this modular characteristic is not clearly reflected in the protein sequence.

7.1 Equilibrium

The order parameter $p(z)$ which is defined in Equation 6.2, as a function of the denaturant factors z, temperature or urea concentration, presents a classical sigmoidal shape, see Figure 6.2. Such curve provide a global information on the protein at the equilibrium and this can be interpreted as a two-state-like protein behaviour. The resulting order parameter shape if found to be even sharper than the experimental data maybe due to the

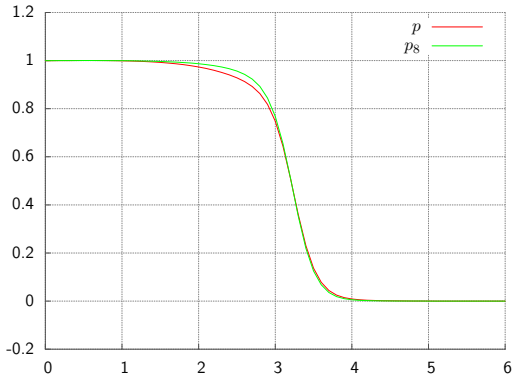


Figure 7.1: Comparison of the signals of the order parameter p, Equation 6.2, and p_8 as functions of the denaturant concentration. These signals are somewhat analogous to the experimental measures of far-UV CD band and of the fluorescent signal from Trp8. Notice that the two signals present great agreement which is usually interpreted in support of two-state behaviour, even if some differences are evident below c_m.

model energy function that tends to enhance cooperativity [1].

Experimentally, folding cooperativity is inferred when the molecule present similar temperature behaviour of different observables, possibly giving local and global information on the protein structure. To mimic the experimental procedure, in Figure 7.1, we compare the global order parameter signal $p(c)$, which can be associated to the far-UV CD band used in Ref. [42, 52], to the order parameter restricted to the Tryptophan on the site 8 ($p_8(c)$), which can be related to the fluorescent signal of Trp8. The latter can be calculated analogously to the global order parameter, substituting the global m with $\langle m_8 \rangle$ and using its baselines. The two curves overlap to a great extent which experimentally is usually interpreted as two-state behaviour.

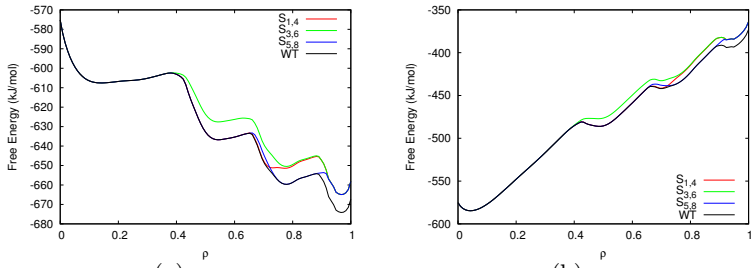
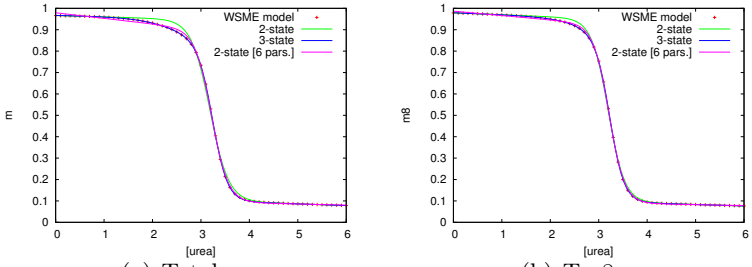


Figure 7.2: Examples of the Myotrophin free-energy profiles as a function of the fraction of native residues $\rho = M/N$. Panel *a*: renaturing conditions (c=0). Panel *b*: denaturing conditions (c=12). Despite the global cooperativity of the folding process found in the order parameter analysis, the free-energy profile present a multi-minima behaviour, suggesting the present of low populated intermediate states. With the wild type curve (black) the behaviour of destabilizing mutation are reported. Mutation destabilize respectively N-terminal region (red line), central region (green line) and C-terminal region (blue line).

7.1.1 Multi-minima Free-energy Profile

On the other hand, while analysing the free-energy profiles, as a function of the fraction of native residues $\rho = \frac{M}{N}$ one observe four minima, in both strongly renaturing and denaturating conditions as shown in Figure 7.2.

The protein two-state-like behaviour that implies cooperative folding seem to be in contradiction with the resulting multi-minima energy profile, for which one can expect the presence of intermediates. In the wild type species, where this behaviour is found, this apparent puzzle is solved by observing that for most temperatures or denaturant concentrations, the intermediate minima are found at a free-energy higher than the native or unfolded minima, and almost never become sufficiently populated to affect the two-state effective behaviour of $p(z)$.



(a) Total (b) Trp8

Figure 7.3: Fits of the fraction of native residues m with a three-state model (blue line), with a two-state model and two parameters (green), and with a two-state model and 6 parameters (purple) in the case of chemical denaturation. While the two-state model provide a good fits of the WSME data, the three-state model results in a better description of the fitted data. However, notice that the 6-parameters, two-state fit could be good enough to stay within the experimental error, in the case of experimental data points. The same analysis has been applied to the entire molecule (a) or restricting the attention to the Tryptophan in the eighth position (b) where the $m8 = \langle m_8 \rangle$ is reported. In the latter case the same observation as in the entire-molecule case can be done.

Actually, a deeper analysis of the free-energy profiles reveals that a intermediate minimum became relevant and significantly populated for a small range of the denaturant factor z close to the transition. Experimentally the folding process is characterized by the estimation of the structure formation both globally, such as far and near UV circular dichroism involving the stabilization of the secondary and tertiary structures respectively, and locally, such as Tryptophan fluorescence. In their work Lowe and Itzhaki [42] show that for Myotrophin all the structure formation measures indicate a two-state-like folding. Figures 7.3(a) and 7.3(b) show a picture of m and $\langle m_8 \rangle$ as function of the denaturant concentration. These quantities provide, respectively, a global information on the formation of folded residues, and a local one on the neighborhood of Trp8, and should roughly correlate with the CD and the fluorescence experimental results. A detailed analysis reveals that the three-state model reproduces the WSME model data with higher accuracy than the two-state one. On the other hand the two-state model, and in particular the one model fitted with 6 parameters that account also for the baselines, while not in perfect agreement with the simulation data, is close enough to be considered within the experimental errors.

Effects of Mutations Mutations perturb the behaviour of the folding process destabilizing certain atomic contacts localized on a particular region of the contact map. Figure 7.4 show the order parameter response to the perturbation of the energy function localized in different regions. A destabilization of central regions affecting helices 3 to 6 ($S_{3,6}$) lowers the mid-transition concentration and basically preserve the cooperativity of the folding process represented by the shape of the sigmoidal curve. Mutations on N-term ($S_{1,4}$) and on C-term ($S_{5,8}$) enhance the role of the intermediate reducing the overall cooperativity of the protein, and can induce a plateau as shown in Fig. 7.4.

The stability of every residue i in the wild type molecule against denaturation can be accounted for by the average $\langle m_i \rangle$. Figure 7.5 show this

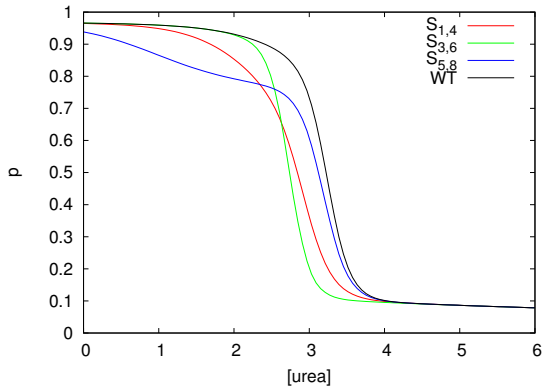


Figure 7.4: Chemical unfolding of Myotrophin due to urea denaturing action. We report the behaviour of the wild type molecule and three mutation (the same as in figure 7.2). Mutation on central region ($S_{3,6}$) seem to only shift the curve, while mutations involving external residues also affect the overall cooperativity (being the C-terminal $S_{5,8}$ mutation the most effective).

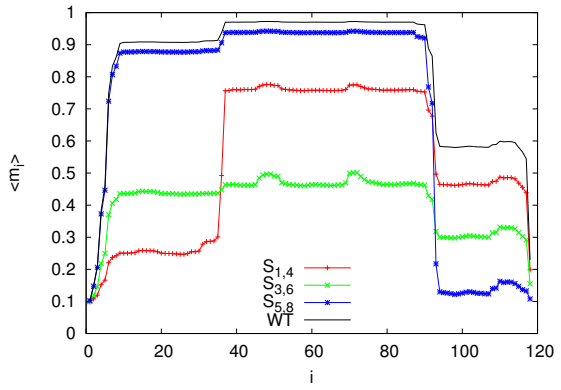
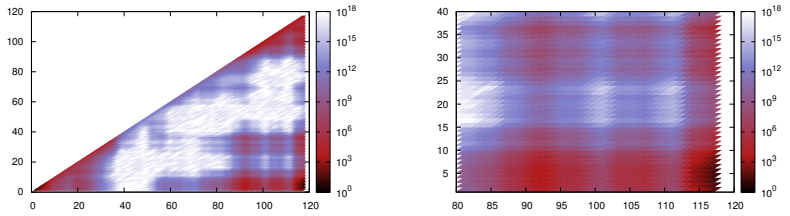


Figure 7.5: Profile of $\langle m_i \rangle$ in the WT and mutated cases, for a denaturant concentration slightly lower than the mid-folding concentration ($T = 2.8$). It represents the equilibrium probability of each residue m_i to be in the *native* state. The C-term repeat is the less stable for the WT molecule. While each mutation concerning external residues ($S_{1,4}$ and $S_{5,8}$) decreases the related-repeat stability, the central mutation $S_{3,6}$, acts on the entire molecule stability.

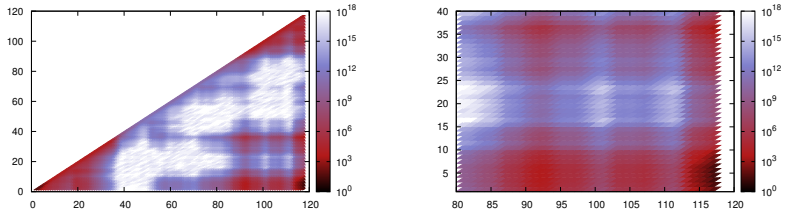
stability for each residue of the Myotrophin, for the wild type (WT) and the three mutation analysed before. The study is performed at $c = 2.8$M, slightly below the mid-concentration in order to enhance and expose the difference of behaviours. The profile of $\langle m_i \rangle$, in the wild type case, shows a clear difference of the central region "nativeness" with the behaviour of the external helices, displaying how the central region of the molecule is more stable compared to external helices, in particular the C-terminal ankyrin repeat is sensibly less stable of the rest of the protein. Mutations introduce an instability into the regions affected by the perturbation, in particular in the external ones, $S_{1,4}$ and $S_{5,8}$, where the one affecting the N-term invert the stability of the terminal repeats. The mutation $S_{3,6}$, affecting the central region of the polymer, induces an overall diminution of the molecule stability decreasing the difference in stability of the different repeats.

These findings are reflected also in the profiles for the same three mutants reported in Fig. 7.2, that can now be more easily interpreted. In this case the free-energy profiles are affected by the perturbation introduced and depart from the wild type behaviour at a certain value of the fraction of native residues ρ both in the folding and in the denaturing conditions. The three perturbation present the same profile in the unfolded minimum, at low values of ρ, as the wild type protein, suggesting that the contacts affected by the mutation are not formed. The first mutant, whose free-energy profile departs from the general behaviour, is the one involving the central region of the molecule, suggesting that this region contributes the most to the free-energy profile at low values of the reaction coordinate. This suggest that contacts localized on the central part of the molecule are involved in early formation (late denaturation) of low structured states. Moreover, the profile at intermediate value of the reaction coordinate is substantially parallel to that of the wild type: these configurations are evenly perturbed by the mutation. On the other hand, the free-energy profile of mutations affecting external regions, as $S_{1,4}$ and $S_{5,8}$, depart from the wild type behaviour at higher values of the reaction coordinate, indicating that they affect the formation (denaturation) of higher structured states, closer to the native state. In particular C-term-affecting mutation, change the free-energy profile only of the native state, suggesting that the helices involved reach the folded state when the rest of the molecule is already arranged in a structured state.

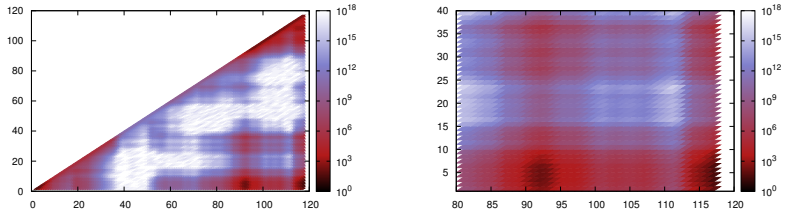
This description seems to suggest a particular folding pathway for the Myotrophin folding and unfolding processes, although we have to be careful while taking any conclusion over the kinetics based on the equilibrium analysis. Effects and consequences over the folding pathways and out-of-equilibrium kinetics will be explained in the following sections.



(a) WT (b)



(c) S_3 (d)



(e) S_7 (f)

Figure 7.6: Inverse native strings probabilities $\frac{1}{\nu_{i,j}}$ (Equation 6.7), at folding conditions $c = 0$, for the wild type and two mutants. On the right: detail of the bottom right corner of the maps, corresponding to the region close to the native structure. These equilibrium maps cannot be directly related to the folding kinetics, yet the qualitative folding mechanisms that can be read out from their analysis agree with the quantitative data from the MC kinetics (see Sec. 7.2 below). Indeed, the region around position (30,92) corresponds to the formation of a central nucleus encompassing the two central ankyrin repeats. From there, the native state at the bottom right can be reached either elongating rightwards and then downwards (C-term, a_f pathway) or downwards and then rightwards (N-term, b_f pathway). Notice that along both pathways after the formation of the central nucleus, there are two regions of low probability, suggesting early and late barriers on both pathways. It is possible to guess that the

7.1.2 The structure of the Intermediate Minima

The analysis of the free-energy landscape goes through the projection of the total landscape over a "reaction coordinate", giving us a global picture of the system. This, however, is a risky operation as this simplification can hide important features of the original free-energy landscape, and the election of a "good coordinate" is generally a non trivial challenge. A unidimensional reaction coordinate may hide the presence in the global landscape of heterogeneous pathways towards the energy minimum.

A complementary free-energy landscape projection has to be introduced in order to analyse the possibility of the pathway heterogeneity proposed by Lowe and Itzhaki [41]. The two-dimensional picture of the inverse of the probability, $\frac{1}{\nu_{i,j}}$, as described in Equation 6.7, can give an insight on the free-energy landscape, as reported in Figure 7.6.

It is easy to identify the native spot at the bottom right corner, and the isolated short structures represented by the short strings close to the diagonal, mainly accumulated at the external regions which, as less stable than the central part, are more likely to fluctuate, hence temporarily destroy the formed structure. In addition to those, five extra spots of intermediate structure can be singled out: the central ones (roughly centred at (32,92) and (32,106)) corresponds to the first intermediate of the free-energy profile in Fig. 7.2, that correspond to $\rho \approx 0.5$, while the others, displaced towards the N-term (the spots around (5,92) and (5,106)) or C-term (the region centred at (32,115)1) respectively, are represented by the intermediate around $\rho = 0.75$ in Fig. 7.2. Interestingly, mutants involving contacts at the N-term or C-term present different probabilities at the intermediate spots, and in the regions connecting them, suggesting that also the pathways could be different between the different species.

However, $\nu_{i,j}$ describes the probability of a residue string from residue i to j to be in the native state capped by unstructured regions. While its inverse is correlated to the free-energy landscape, it is important to notice

that for each point (i, j) in this two-dimensional picture the value of $\nu_{i,j}$ integrates-out the states of residues outside the string from $i-1$ to $j+1$ so that it does not correspond to a single state and hence two strings (i, j) and (k, l) with $i < j < k < l$ are uncorrelated by construction: if the two regions appear with high probability, it does not imply that the configuration with both structured regions is especially likely. So, even if these two-dimensional profiles already suggest possible pathways and folding mechanisms, they do not allow a quantitative characterization of the kinetics, and a detailed study of the latter must be performed independently, as in the following section.

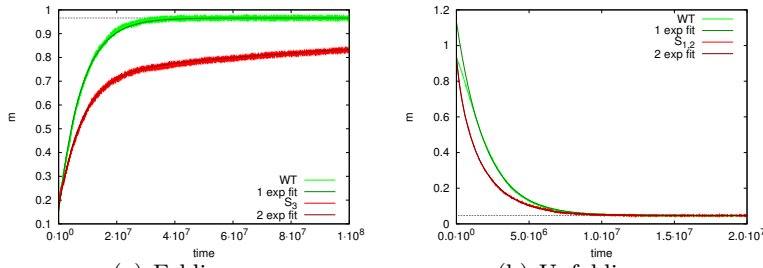
7.2 Kinetics

Both folding and unfolding relaxations outside the equilibrium are studied through Monte Carlo simulation of an ensemble of 2000 molecules in the wild type case and for certain mutations.

7.2.1 Two-state Behaviour

The simulations performed on the wild type species reveal a single-exponential kinetics, as can be seen in Fig. 7.7, which is in agreement with the results in [42]. This simple behaviour is apparently at odds with the multiple minima landscape reported in the free-energy profiles analysed before in subsections 7.1.1 and 7.1.2: to gain some detailed insight on how these characteristics can be simultaneously present, we study the relaxation events of individual molecules.

Single Molecule Simulations. Some representative examples of single-molecule relaxations for the wild type species are reported in Fig. 7.8, for the folding and unfolding case.



(a) Folding (b) Unfolding

Figure 7.7: Plot of the fraction of native residues $m(t)$ as a function of time, for the folding *(a)* and the unfolding *(b)* process, for the wild-type and for two mutants. The wild type protein shows a single-exponential behaviour, also common to the majority of the mutants. The two mutants here are chosen to represent the two-exponential behaviour, for comparison. The average value was calculated over 2000 molecules simulations.

We have seen that, neglecting the ubiquitous structure fluctuations, it is possible to identify some precise patterns in the relaxation process for both folding and unfolding processes, summed up in the figure. Typically, in the folding relaxation, the molecule follow one of two main behaviours: in a early stage, the stabilization of the central nucleus involve all the molecules with the formation of stable structure in the four central helices (ankyrin repeats two and three). The formation of structure at the interface between the second and third repeats typically triggers the immediate stabilization of both of them (even though this might be an artefact of the model, that just considers interactions if they take place within a native string). Once the central structure is formed, the polymer face a crossroads given by the eventual stabilization of the C-term (Fig. 7.8(a)), or the stabilization of the N-term (Fig. 7.8(b)). Finally, after a variable transient, the molecule reach the completely native state.

In the unfolding relaxation case, the process follows a similar but not completely symmetrical trajectory: the molecule from the native state, un-

fold first an external ankyrin repeat (N-term in Fig. 7.8(c) and C-term in Fig. 7.8(d)), then the unstructured region expand one repeat toward the central part and in the final step reach the completely unfolded state.

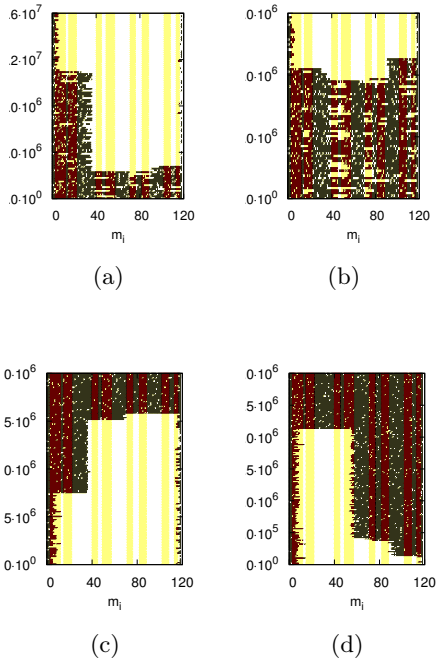


Figure 7.8: Single-molecules simulations of folding (panels a,b) and unfolding (c,d) events. Panels a,c: relaxation along pathway "a"; panels b,d: relaxation along pathway "b" (see text for details). The residue state $m_i(t)$ is reported with different colours: black if unfolded ($m_i = 0$) and white if native ($m_i = 1$). The vertical stripes indicate the positions of the α-helices.

Characterization of Pathways. To characterize in a more quantitative way these behaviours, we have identified, for each single molecule trajectory, the rate limiting steps, by inspection of the time differences Δt between stabilizations of different strings $t^{(x)}_{\alpha,\beta}$, where $x = f, u$ for folding and unfolding trajectories respectively, and identification of the biggest Δt between consecutive stabilizations. In order to classify folding and unfolding pathways we identify the formation (disruption) of some key strings that trigger the successive events towards the N- or C-term. In both folding and unfolding case, we can distinguish two pathways, that we call a_x, b_x where $x = f, u$ for the folding and unfolding case. Pathway a_x is characterized by the C-terminal part getting structured earlier in the folding process, and disrupting later in the unfolding one. On the contrary the b_x pathway favours earlier stabilization of the N-terminal part in the folding process, and its longer persistence in the unfolding one. However, folding and unfolding pathway of the same kind do not coincide, so that we distinguish them with the f, u labels. The detailed definition is as follows: for the folding process, we find that the event triggering the a_f pathway is the formation of a native string encompassing helix 4 to 7 before that of a native string from helix 2 to 5; the opposite order characterizes the b_f pathway. For the unfolding process, a_u is characterized by the contacts between helices 6 and 7 lasting more than those between helix 2 and 3, while the order is reverted in b_u. We have seen that with the above definitions, it is possible to classify clearly and uniquely all the single-molecule relaxations (of the wild type and of the mutated species, see below) as belonging to either pathway.

These results are summarized in Table 7.1, where the rate and amplitude for the one-exponential fit and the fraction of molecules through the a and b pathways are reported. We find a dominance of the pathway a_f through the C-term over the b_f, and a slight dominance of b_u over a_u, pointing out that there may be some differences in the topography of the energy landscape in folding and unfolding conditions.

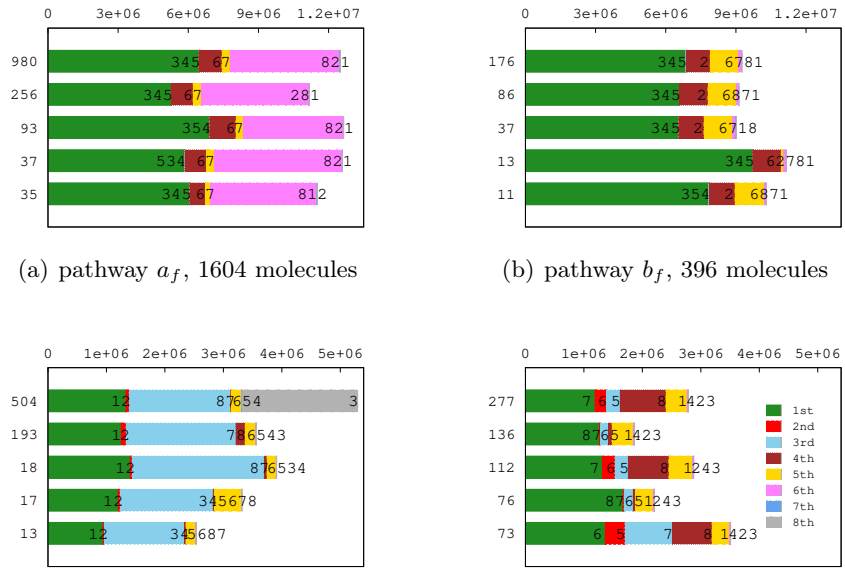
How is it possible that two different pathways are present, while the

	k_1 (10^{-7})	$\lvert c_1 \rvert$	a (%)	b (%)
folding	1.279(9)	0.814(4)	80.2	19.8
unfolding	4.76(8)	0.965(4)	40.2	59.8

Table 7.1: Rates and amplitudes (from 1-exponential fits) and fraction of molecules that select folding pathway a and b, in the case of folding and unfolding relaxations. The errors have been estimated by dividing the proteins in 10 groups of 200 molecules each, evaluating a rate and amplitude from the fit of the average signal of each group, and calculating the mean and deviation of the mean of the resulting population of rates and amplitudes.

folding and unfolding appear as two-state processes? The difference in the fraction of molecules following either pathway suggests that there is a little difference in the free-energy barrier that they have to surmount. This difference cannot be huge, since in that case it would result in rates along each pathway differing by order of magnitudes, which in turn would imply fluxes by just one channel. Moreover, the fact that several minima, connected by different barriers, are found in the free-energy profiles, but the relaxation kinetics is simply exponential (two-exponential fits fail to produce reliable results due to over-fitting, data not shown), implies that either the rate limiting step is represented by crossing the first barrier along the pathway, effectively masking the other jumps, or the different barriers are associated to very similar rates.

This picture is confirmed by the analysis of the average times of helix stabilization (or destabilization, in the unfolding process), reported in Fig. 7.9, where the most representative patterns of secondary structure formation are reported separately for both the folding and unfolding pathways. It is clear from the top panels that the folding pathways are characterized by the formation of helices 3, 4, 5 basically altogether, around t=$6 \cdot 10^6$, followed by the extension, in another million of time steps, toward helices



(a) pathway a_f, 1604 molecules

(b) pathway b_f, 396 molecules

(c) pathway a_u, 804 molecules

(d) pathway b_u, 1196 molecules

Figure 7.9: Patterns of helices stabilization in the folding process (top) and of their disruption in the unfolding one (bottom), from the analysis of 2000 single molecule relaxations, both for folding and unfolding events. Just the five most representative patterns are reported for each folding or unfolding pathway. In the top panels, the horizontal axis represents $\langle t^{(f)}_{\alpha,\alpha} \rangle$, that is, the last time (in units of MC steps) that the helix α turns completely folded in the simulation. In the bottom panel, the corresponding quantity $\langle t^{(u)}_{\alpha,\alpha} \rangle$ for unfolding is reported. The averages are performed on all the molecules n_s following the same succession of events s, which can be read in the labels of the bars; the number n_s is reported at the left of the y-axis. The colour code is the same for all panels, and refers to the order of helix stabilization (destabilization in the unfolding case). Missing colours, as well as grouping of the corresponding labels, indicate that a group of helices folds (or unfolds) almost at the same time. When this grouping takes places, it is quite common to find patterns that differ just for the permutation of grouped labels: e.g. this is the case for the a_f pathways reported in panel (a), where all the patterns are small variation of the same scheme in four steps: folding of helices 3, 4, 5 basically at the same time, then stabilization of helix 6 and then 7 after a short time, and finally completion of the folding.

6 and 7 in pathway a_f, or helix 2 in pathway b_f. Then, the rest of the structure folds almost at once. In the folding process, the longest time is associated to the formation of the initial nucleus, and the second longest one (and close to the former), to the completion of the folding along pathway a_f.

The unfolding process presents as well some common schemes: in the dominant b_u pathway, unfolding proceeds from the C-term (helices 7, 6 and 5), passes through the last and first helix, and finally affects the rest of the N-terminal part. The a_u pathway presents more variability, but the dominant mechanism is given by the a_f pathway covered in the opposite direction, and ending with the central group of helices 3, 4 and 5. Notice that in both the unfolding pathways, the second repeat appears as the last to unfold, and the longest time is usually associated to the unfolding of the last repeat, and in particular of helix 7.

Fig. 7.9 suggests a clear picture of how the folding and unfolding proceed along the a_x and b_x pathways and gives an idea of what are the rate limiting steps, even if it does not inform on the detailed structure of the transition states and nuclei. In fact the latter could contain partially structure helices and hence need not coincide with a collection of fully formed helices, while the times $t^{(f)}_{\alpha,\alpha}$, $t^{(u)}_{\alpha,\alpha}$ inform on when the helix α becomes stably structured or unstructured as a whole, but they are affected by structural fluctuations within the helix. Moreover, averaging the times $t^{(f,u)}_{\alpha,\alpha}$ (that may vary a lot from molecule to molecule) gives no information about the time evolution of any observable.

For these reasons, the picture coming from Fig. 7.9 must be checked and confirmed by further analysis. To this end, we have simulated the effect of mutations, as explained in Section 6.1, and performed the same analysis as for the wild type species.

	folding				unfolding			
	k_1 (10^{-7})	k_2 (10^{-7})	a_f (%)	b_f (%)	k_1 (10^{-7})	k_2 (10^{-7})	a_u (%)	b_u (%)
WT	1.29		80.2	19.8	5.08		40.2	59.8
$S_{1,2}$	1.22		85.1	14.9	5.20	32.0	85.7	14.3
S_3	0.058	1.26	98.7	1.3	5.00		39.6	60.4
$S_{3,4}$	0.116		85.9	14.1	6.74		38.8	61.2
$S_{5,6}$	1.22		80.2	19.8	4.92		41.0	59.0
S_7	1.18		45.3	54.7	6.45		29.9	70.1
$S_{7,8}$	1.26		61.7	38.3	6.95	47.8	1.7	98.3
$S_{1,4}$	1.28		80.5	19.5	5.02	21.4	88.2	11.8
$S_{3,6}$	1.25		78.2	21.8	5.13		39.1	60.9
$S_{5,8}$	1.34		69.1	30.9	6.99	50.9	1.6	98.4

Table 7.2: Rate and fraction of molecules that select folding pathway a and b in the case of folding and unfolding dynamics for different types of perturbation, as compared to the wild type (WT). Here the rates are calculated by fits on the whole ensemble of 2000 proteins. Missing entries in column k_2 mean that the 1-exponential fit was sufficient, and the 2-exponential one would produce over-fitting.

7.2.2 Simulated Mutants and Pathway Heterogeneity

Table 7.2 summarizes the results for the kinetics of the mutants: most of the times, the $m(t)$ signal can adequately be fitted with just one exponential (with rate k_1) both in the folding and unfolding case, while some mutants present a second, faster phase k_2, especially in the unfolding case.

Folding Kinetics

We see that in the folding case, mutations affecting helices 2,3 and 7,8 do not change the rate much, but they are those that tune the flow along the

two pathways. Mutations at the C-term cause the biggest pathway shifts, in agreement with the experimental results [41]. Mutations in the central nucleus (e.g. $S_{3,4}$) affect the rate , but do not change much the distribution along the two pathways.

Mutants $S_{5,6}$, $S_{3,6}$ are associated to contacts that do not stay at a barrier top: their destabilization weakly affects the rates and pathways.

The above results confirm that the formation of the central nucleus of three helices 3,4 and 5 is the rate limiting step for the folding process, followed by a growth of the nucleus towards the C or N term (pathways a_f and b_f respectively).

The barriers associated to these pathways are smaller enough that most of the perturbations that shift the flow cannot "promote" these barriers to be the highest one. As a result, the above sequence of events is preserved in the mutants, and the folding rates are little affected and quite similar to the WT one, while the distribution of the flow changes according to which of the pathways is destabilized. The behaviour of $S_{3,4}$ is consistent with the above observations: here the central nucleus is destabilized, which results in a slower rate with a moderate change in the flows. The only exception to this picture is mutation S_3, that affects all the contacts (local and non-local) of residue 32, located in the loop between the first and second repeat. The destabilization of these contacts leaves the rate for the formation of the central structure unchanged, but produces a second, slower rate, corresponding to the last steps of the folding along pathway a_f, of stabilization of the first repeat structure. Accordingly, the corresponding folding flow goes almost completely through the a_f pathway.

Consistent with this interpretation, $S_{1,4}$ does not involve relevant changes in either the folding rate or fluxes, and appears downhill with respect to the crossing of the last barrier along the a_f pathway.

Unfolding kinetics.

As in the folding, mutations affecting the external regions (first and last ankyrin repeats) cause the biggest changes in the flow through the pathways; these changes agree with those observed in the folding case: a mutation causing a larger flux towards the a_f pathway in folding will also cause an increase in the a_u flux in unfolding, though of different magnitude. Moreover, these mutations are accompanied by the appearance of a second, faster rate, signalling that a part of the structure unfolds before the rate limiting step; the latter rapidly leads to the completion of the unfolding.

On the other hand, mutations affecting the central region do not cause major changes in either the rate or the flux distribution along the two pathways, with respect to the wild type. Interestingly, in the S_3, $S_{3,4}$ and $S_{5,6}$ species the changes in the flux have opposite sign in the folding and unfolding processes, again suggesting that the choice of the pathway is not controlled by the central repeats.

As Table 7.2 suggests, the presence of two rates, in either folding or unfolding, is not related to the two different pathways for the kinetics, as one could naively think at the beginning. This is even clearer in Table 7.3, where the single or two-exponential fits are performed separately on the subsets of molecules following either pathways: in general, the need for a two-exponential fit for the full ensemble is associated to the presence of two rates in either the a or the b pathway.

Another interesting thing is that the rates along the less populated pathways are usually comparable or greater than those along the corresponding dominant one. This apparent contradiction can be explained if one considers that, due to the restriction of the fit to a specific subset of the molecules, the resulting rate is not related to the equilibrium distribution, and it gives no information on the height of the free-energy barrier along the pathway. This can be easily understood by considering, for example, wild type molecules: after the formation of the nucleus of the two central repeats, they have to

	a_f		b_f		a_u	b_u	
	k_1 (10^{-7})	k_2 (10^{-7})	k_1 (10^{-7})	k_2 (10^{-7})	k_1 (10^{-7})	k_1 (10^{-7})	k_2 (10^{-7})
WT	1.27		1.36		4.53	5.93	
$S_{1,2}$	1.19		1.40		5.14	27.8	
S_3	0.056	1.24	0.53	3.68	4.10	5.76	
$S_{3,4}$	0.115		0.109		6.74	6.73	
$S_{5,6}$	1.18		1.34		3.87	5.38	
S_7	1.13		1.18		7.11	6.40	
$S_{7,8}$	1.25		1.24		37.2	6.86	48.2
$S_{1,4}$	1.25		1.42		5.06	23.2	
$S_{3,6}$	1.21		1.37		4.51	5.73	
$S_{5,8}$	1.30		1.35		49.2	7.12	54.2

Table 7.3: Rates calculated on the subsets of molecules following the different pathways of folding and unfolding, as found in Table 7.2.

choose whether to follow the a_f pathway, with a low barrier, or the b_f one, with a higher one. Table 7.2 shows that the majority of them will follow the former. Therefore, the fraction of molecules along the b_f pathway will be given by those that choose that pathway early, i.e. in times shorter than the typical folding time along a_f, producing an apparently faster rate. This can also be seen in the top panels of Fig. 7.9: the formation of helix 2 in the b_f pathway takes place in a time of the same magnitude as the selection of helix 6 and 7 in the a_f one. After the formation of helix 2, the folding along b_f is faster than the competing one. Thus, a fit restricted to the b_f ensemble yields naturally a faster rate than the one found for the a_f pathway.

Indeed, we have checked that if the b_f pathway is imposed to the wild type protein, by associating an energy penalty to the formation of long native string towards the C-term, so that the protein cannot "escape" through the a_f pathway, the resulting rate $k_{b_f} = 1.20 \cdot 10^{-7}$ is lower than that

mut	folding		unfolding	
	a_f %	b_f %	a_u %	b_u %
WT	80.2	19.8	40.2	59.8
E17V/D20L	31.8	68.2	11.6	88.4
A9G	82.5	17.5	83.1	16.9
A115G	71.9	28.1	25.7	74.3
A115G/A9G	78.3	21.7	70.6	29.4
A115G/E17V/D20L	24.5	75.5	7.3	92.7
A9G/E17V/D20L/A115G	38.9	61.1	42.6	57.4

Table 7.4: Fraction of molecules following pathways a, b in both folding and unfolding kinetics. Mutations E17V/D20L are stabilizing, while the others are destabilizing. As expected, E17V/D20L, stabilizing the N-term, shifts both the folding and unfolding rates towards the b pathway, while A9G and A115G, destabilizing the N and C-term respectively, shift towards the a and b pathway. The combined effect of the various mutations is the expected one, in terms of mutual enhancement or suppression, according to whether the mutations shift in the same or opposite directions.

observed in free wild type.

Multiple mutations

In a recent work, Lowe and Itzhaki [41] induce switches between the pathways by engineering multiple mutants. A number of single site mutants are compared and then combined to analyse the effects of those perturbations to the system pathway. In order to compare our model to those results, we have simulated the effect of those mutations applying a stabilizing or destabilizing perturbation, as explained in Section 6.1.

The model predictions for such mutants, reported in Tables 7.4 and 7.5

mut	folding k 10^{-7}	unfolding k1 10^{-7}	unfolding k2 10^{-7}
WT	1.29	5.08	
E17V/D20L	1.59	2.90	
A9G	1.20	4.83	15.48
A115G	1.23	6.52	
A115G/A9G	1.15	10.5	
A115G/E17V/D20L	1.55	2.59	7.83
A9G/E17V/D20L/A115G	1.47	10.2	

Table 7.5: Rates of the mutations in both folding and unfolding kinetics. In general, the folding rates are weakly increased or decreased, and vice versa for the unfolding rates, according to whether the mutation is stabilizing or destabilizing. The shifts agree qualitatively but not quantitatively, with the experimental results.

show that the redistribution of the flux along the pathway upon combination of point mutations is qualitatively as expected from the discussion in the previous section, so that a mutation favouring pathway a will balance the effects of a mutation favouring pathway b, recovering at least partially the WT flows, even if with smaller folding and bigger unfolding rates. It must be noticed, though, that the ratio between the corresponding rates do not reflect the experimental results, and the model fails to give quantitative predictions of the rates.

7.3 Discussion

The results reported show that the WSME model reproduces qualitatively the experimental behaviour: indeed, we find a two-state-like equilibrium, a two-state-like kinetics with transient intermediates, and also two pathways

in kinetics, characterized by the order of structure formation at the C- and the N-term. Moreover, we find that precisely targeted perturbations of the contact interactions at different positions along the chain allow to induce pathway switches, as seen in experiments, or the stabilization of some transient intermediate resulting in a faster phase, and provide a way to probe the folding mechanisms to a great detail. The intrinsic complexity of the free-energy landscape of the protein Myotrophin is evident already from the analysis of the free-energy profiles, Fig. 7.2. However, in this case such 1-dimensional projection is not sufficient to suggest the details of the kinetics, since the two different pathways cannot be distinguished just on the basis of the number of native residues, which is the natural reaction coordinate of the model. Moreover, it is important to notice that the free-energy profiles present small barriers (less than 3 RT), which seems at odds with the much slower rates found in the simulations. We see three possible reasons for such difference: first, in the presence of two different pathways in the configuration space, but with barriers located at similar values of the reaction coordinate and thus roughly overlapping in the projection (see Fig. 7.6 and the relative discussion), the profile will be always more representative of the lower of the two overlapping barriers, since, by construction, it is more representative of the states with higher Boltzmann weight. However, if, as in this case, the lower early barrier is on the a pathway and the lower late barrier is on the b pathway, the true barrier on each pathway will be higher than it may be inferred from the profiles. Second, the presence of wide and roughly flat regions between each minima and the following barrier slows down the rate, according to to the role of the curvature of the profile in Kramers' theory. Third, the projection collects together, at the same values of the reaction coordinate, configurations that can be highly different (in terms of the Hamming distance), especially in the unfolded region. This is irrelevant to kinetics as long as the motions in the transverse directions are fast enough to be relaxed at equilibrium when considering displacements along the reaction coordinate, but this is not necessarily granted for proteins with pathway heterogeneity. Independently from which of the above

possibilities is more relevant, the important message from the above observations is that any quantitative conclusion about kinetics, derived from the analysis of the free-energy profiles, must be drawn with care, especially for proteins with a pathway heterogeneity.

Yet, some hints for such a read-out of the kinetics come from the analysis of the equilibrium probability for the formation of native strings, Fig. 7.6: even if such probabilities do not constitute a free-energy map, and cannot be used to predict the kinetics by transition state theory or by solving a diffusion equation, they suggest the important spots playing a key role in the intermediate states, the possible folding pathways, and the way mutations can affect the kinetics, redirecting the folding and unfolding fluxes.

The picture that emerges, and that could probably be generalized to other repeat proteins, is that in such proteins multi-minima free-energy profiles are the rule, with intermediate states related to the completion of the folding of whole repeats or substructures of them. In this framework, the cooperativity in the equilibrium unfolding reported in experiments would be attained by a "designed" free-energy landscape, such that in all conditions the intermediates are associated to free-energies substantially higher than those of the native and unfolded states. Such design would involve sequence-heterogeneity between the different repeats, to ensure different degrees of stability to different partially folded structures. Mutations can alter this situation[40], and indeed we see that cooperativity is reduced by perturbing the N-term, and especially the C-term of Myotrophin. Two-state kinetics would most likely emerge, in such a multi-minima landscape, when the rate-limiting step coincides with the crossing of the first barrier encountered in the folding or unfolding process, and masks the crossing of the following barriers. Again, mutations may "promote" different barriers to the status of rate-limiting step, thus involving multi-state kinetics and/or pathway heterogeneity.

Unfortunately, the predictions of the model, especially for the kinetics, cannot be made quantitative, at least at this level of simplicity. In the

model we adjust only two parameters to reproduce the temperature and concentration of the folding midpoint, and at this level of simplification, we cannot reproduce the ratio of the rates between the mutants studied in the experiments. Also, the central nucleus that we find in the folding process is not reported in the interpretation of the experimental data by Lowe and Itzhaki (that propose a three-state dominant pathway plus a two-state secondary one to interpret their data, and assume that the total relaxation rate is just the sum of the rate along the two pathways). The enhancement of the role of such a common central nucleus, that delays the choice of either folding pathway to higher values of the reaction coordinate, might therefore be a model artefact, due to the model feature of considering the interaction energies as just proportional to the number of contacts of each residue: this may effectively penalize the terminal helices, that make fewer contacts.

However, it is important to stress that the model gives predictions which go beyond the experimental results, as for instance the detailed information about the pathways, providing useful conceptual frameworks for the interpretation of the experimental data. An important suggestion coming from the model predictions is that folding and unfolding pathways are not necessarily the same pathway, covered in opposite directions: the different denaturant concentrations (or temperature conditions) may involve subtle but important changes in the energy landscape, so that the overall mechanisms (for instance, the two-state, two-pathway kinetics) do not change, but the details of the pathways do. Indeed, in the presence of two possible pathways and a strong bias towards folding (or unfolding), it is most likely that at any "fork" in the pathway the protein will follow the trail with the smaller barrier at that point, and will be stuck on it, since backwards jumps will be strongly suppressed. In the present case, the lower folding barrier at lower values of the reaction coordinates on the a_f pathway, and the lower unfolding barrier at higher values of the reaction coordinates on the b_u pathway, produce heterogeneity in the folding and unfolding pathways. Independently on how realistically this mechanism describes the behaviour of Myotrophin, it is an

important warning for the design of simple interpretation frameworks for experimental results: if on the one hand, a phenomenological model must be kept as simple as possible, to avoid over-fitting and the introduction of too many parameters, on the other hand, the use of simple models as the present one, with very few free parameters, may represent a useful alternative to get a grasp on the key mechanisms of the folding and unfolding process.

Conclusions

To conclude we would like to summarize the most important findings exposed in the two part of this Thesis, as well as to present some perspectives that arise from the results. This Thesis deals with the application of analytical and computational methods, typical of statistical mechanics, to biological problems, and shows that, if the latter can be successfully mapped onto a statistical-mechanical system, the powerful tools of this field can be applied to the study of biologically relevant systems yielding qualitative and even quantitative predictions.

The first, and sometimes most important step in this approach, is to find the best mapping of the original problem on a physical system whose energy-function describes all the relevant interactions and constraints of the original model, and at the same time is simple enough to be conveniently dealt with by analytical or numerical methods.

In this Thesis we have applied this approach to two apparently different biological problem: in the first we have approached the problem of peptide sequencing from Tandem Mass Spectrometry, while in the second we have analysed the folding behaviour of the small repeat protein Myotrophin.

These problems have been mapped to a unidimensional statistical system with interactions that are either local, with only nearest neighbours interaction, or that can be reduced to block interactions: in both cases,

they are amenable to an exact calculation of the equilibrium distribution through the application of the transfer-matrix approach.

Protein Sequencing In the first part we have introduced the problem of peptide sequencing by Tandem Mass Spectrometry, which represents an important challenge in Proteomics. Tandem Mass Spectrometry is a common, fast and reliable technique that provides a mass spectrum fingerprint of a peptide, containing the necessary information to infer the residue sequence of the target peptide. Tandem Mass Spectrometry, thanks to his simplicity and relative low cost, is vastly used, generally embedded in an automated high-throughput pipeline. This lead to a huge amount of data to be recollected and to the need for a machine-driven tool for their interpretation, a task complicated by the presence of noise or missing peaks.

The present approach to the interpretation problem relies mainly on database-searching algorithms, that are the most effective in "standard situations", when the unknown protein sequence is present in the protein sequence database or proceeds from an already sequenced genome. However, these algorithms become inefficient or even confusing if used in non-standard situations when there is no previous database knowledge of the protein, or the latter has undergone important mutations or post-translational modifications. Alternatively, *de-novo* algorithms do no suffer from these limitations, but their search space is much bigger, and are more easily fooled by noisy or missing peaks, so that the quality of their predictions is usually much lower. Finally, the problem of the assessment of the quality of the predictions is an important and substantially unsolved problem, common to both strategies.

Our *de-novo* approach in dealing with this challenging problem relies, on one hand, on the mapping of the sequencing problem on the the study of the equilibrium behaviour of a physical system, for which we design a suitable *ad-hoc* potential derived from the analysis of the typical ions distributions in the spectrum space, and, on the other hand, on the exact calculation of

the partition function and other relevant thermodynamic variables of the system, from which the precursor sequence can be inferred.

The resulting method *T-novoMS* introduces a fictitious temperature as a parameter of the algorithm. The latter, as in a real thermodynamic system, behaves as a switch controlling the amount of fluctuations and discerning two main regimes: the low temperature regime where the system is trapped in the energy minimum reflecting the information included in the target spectrum, and the high temperature regime where it explores a bigger part of the configuration space. In this way, at low temperature our algorithm is similar to other *de novo* softwares and produces a prediction of the sequence that best fits the spectrum. At higher temperatures, it explores suboptimal regions of the sequence space, which can be used to asses the validity of the prediction.

The algorithm has been tested against a reliable database of double charged peptide spectra and on the same database have been tested some of the available *de-novo* software (NovoHMM, PepNovo, Lutefisk). Our algorithm produces results comparable with the existing *de-novo* softwares (even if it doesn't reach the performance of the best one), while it exhibits some useful features that lack in the others. The most outstanding feature is the temperature control that, combined with the possibility to calculate exactly the probability distribution, provides the user also with a probability profile that accounts at once for the whole sequences space and gives a temperature-dependent, Boltzmann weight to each sequence. In a future development, this will exploited to perform a peptide-database search as a postprocessing, matching the peptides on the probability profile: hopefully, the correct parent sequence should be the one that "picks" the highest probability from the profile, among all the database sequences. Another feature introduced by this method is the possibility to use thermodynamic functions as a proxy for the quality of the interpretation, reducing false positives: we have seen that the entropy of the system at $T = 1$ correlates with the F-value measuring the quality of the prediction, and that it is possible

to identify some thresholds for the entropy value that can act as a confidence interval, to estimate the probability that the F-value exceeds a given threshold.

The future perspectives include improvements on the final algorithm, mainly in the form of the energy function. Future improvements will probably come from a redefinition of the learning database, with more specific and stringent filters and a more specific definition of the precursor mass, that has been shown to be fundamental to improve model predictions. A separate characterization of the ion distributions for different spectrometers can also improve the definition of the energy function, as the underlying physics of the fragmentation and separation processes may produce different spectrum patterns. An important improvement would probably come from the possibility to adjust the chemical potential (and hence, the resulting length of the predicted peptide) automatically for each spectrum.

Protein Folding In the second part of this Thesis we applied the WSME model for protein folding to describe the interesting behaviour of the repeat protein Myotrophin. This protein is structured in a modular way, where the modules or repeats present an intrinsic stability and are arranged in a linear way. This proteins poses some challenging questions, since it seems to fold in a cooperative way despite its modularity, whereby it has attracted the attention of the researchers.

Our results reproduce qualitatively the experimental behaviour of the molecule. Cooperativity is reproduced both in the equilibrium analysis and in the kinetics, and a deeper investigation show that this is compatible with the multi-minima free-energy landscape predicted by the model. The simulation of the kinetics of the folding process reveals the presence of pathway heterogeneity, as proposed by Lowe and Itzhaki [42] to explain in a unique framework the behaviour of the wild type protein and its mutants. Moreover, the model is applied to the analysis of several mutants corresponding to the application of local an non-local energy perturbations,

and predicts behaviours qualitatively similar to the experimental ones (in particular, it shows that mutations may act as a switch in the flow of the folding/denaturing molecule towards a pathway or the other).

A quantitative modification of the model can be introduced by fitting the model parameters to the available experimental data, as shown by Bruscolini and Naganathan [7]. Parameters are adapted to reproduce the picture of the C_P through the use of a phenomenological expression derived by the WSME model. This model has been shown to describe successfully and quantitatively the folding behaviour of a downhill and a two-state-like proteins[7]. Unfortunately, such refinement of the model could not be applied in the present case, since it is based on the fit of calorimetric data, which are not presently available for Myotrophin.

In this framework an interesting approach could be represented by the study of similar repeat proteins composed by identical modules, the so called consensus repeat proteins. These proteins are composed by the repetition of the same string of residues; they are not natural protein, but have been synthesized in the laboratory and their structures have been resolved. The study of consensus repeat proteins could indeed give an insight into the processes governing the microscopic dynamics in modular polymers.

In conclusion, and according to the above observations, we believe that the use of simple statistical-mechanics models will provide a good wealth of results for biologically inspired problems in the next future.

Conclusiones

Concluyendo, se resumen los resultados más importantes descritos en esta Tesis y se presentan algunas propuestas para estudios futuros. En esta Tesis hemos tratado la aplicación de métodos analíticos y computacionales a problemas de tipo biológico, y hemos mostrado que, si este último puede ser reescrito como un sistema mecánico-estadístico, las poderosas herramientas de este campo pueden ser aplicadas al estudio de sistemas biológicamente relevantes, llevando a predicciones cualitativas y, en algunos casos, hasta cuantitativas.

Siguiendo este enfoque, el primer paso y a veces el más importante, es encontrar la mejor manera de traducir el problema original en un sistema físico cuya función energía describa todas las interacciones relevantes y las ligaduras del problema, y al mismo tiempo dicha función energía sea lo bastante simple como para ser convenientemente tratada con métodos analíticos o numéricos.

En el trabajo que hemos presentado, se ha aplicado esta metodología a dos problemas biológicos aparentemente muy distintos: en el primero de los dos hemos afrontado el problema de la secuenciación de péptidos procedentes de Espectrometría de Masa Tándem, mientras que en el segundo hemos analizado el plegamiento de la Miotrofina, una pequeña proteína de tipo modular.

Estos problemas han sido traducidos a un sistema estadístico unidimensional con interacciones locales a primeros vecinos, o con interacciones que puedan ser reducidas a bloques: en ambos casos se puede calcular exactamente la distribución al equilibrio a través de la aplicación de un calculo de tipo matriz de transferencia.

Secuenciamiento de Proteínas En la primera parte hemos afrontado el problema de la secuenciación de péptidos que representa un desafío importante y aún sin resolver en el campo de la Espectrometría de Masa Tándem. Ésta es una técnica común, rápida y fiable la cual proporciona un espectro de masa de un péptido que contiene la información necesaria para poder inferir la secuencia de los residuos que lo componen. La Espectrometría de Masa Tándem, gracias a su simplicidad y relativamente bajo coste, es ampliamente utilizada y generalmente incrustada en una instrumentación high-throughput automatizada, donde una gran cantidad de péptidos vienen analizados automáticamente. Por este motivo la Espectrometría de Masa Tándem permite recolectar una enorme cantidad de espectros que necesitan una herramienta automatizada para su interpretación.

Los algoritmos disponibles actualmente son efectivos si aplicados a un número restringido de casos, como por ejemplo el análisis de proteínas cuya secuencia es conocida y recolectada en una base de datos, por otra parte éstos se vuelven ineficientes si utilizados en situaciones no estándar como la caracterización del contenido proteico de sistemas cuyo proteoma es todavía desconocido o el secuenciación de proteínas mutadas o modificadas.

Nuestro enfoque frente a este desafío se basa en la traducción del problema de secuenciación al estudio del comportamiento en el equilibrio de un sistema físico. Este estudio se basa en el diseño de un adecuado potencial *ad hoc* derivado del análisis de la distribución típica de los iones en el plano del espectro, y en el cálculo exacto de la función de partición y de otras variables termodinámicas relevantes del sistema, a partir de las cuales se puede calcular la secuencia del precursor.

Nuestro método *T-novoMS* introduce una temperatura ficticia como parámetro del algoritmo. Este parámetro, como en un sistema termodinámico real, actúa de interruptor que controla la cantidad de fluctuaciones presentes en el sistema y distingue dos regímenes: uno a baja temperatura donde el sistema está atrapado en un mínimo energético que refleja la información presente en el espectro estudiado, y otro a alta temperatura donde el sistema puede explorar una parte más grande del espacio de las configuraciones. A baja temperatura, nuestro algoritmo se comporta de manera similar a los demás *de novo* algoritmos prediciendo la secuencia que mejor se ajusta a los datos experimentales. A temperaturas más altas, el sistema explora regiones del espacio de las secuencias alejadas del mínimo energético, que puede ser útil para evaluar la validez de la predicción.

El algoritmo ha sido testado mediante una base de datos de espectros acoplados con una interpretación fiable de la secuencia de los péptidos originarios, todos doblemente cargados, y sobre la misma base de datos han sido testados algunos de los algoritmos *de novo* disponibles (NovoHMM, PepNovo, Lutefisk). Nuestro algoritmo produce resultados comparables con los programas existentes y además exhibe algunas característica útiles relacionadas con el sistema termodinámico asociado que no se encuentra en los demás. Sobresale la posibilidad del algoritmo de controlar la temperatura de trabajo junto con la posibilidad de calcular exactamente la distribución de probabilidad al equilibrio, estas características en su conjunto permiten considerar al mismo tiempo el entero espacio de las secuencias.

Futuros desarrollos incluyen mejoras en el algoritmo final, sobre todo en la forma de la función coste. Mejoras en el algoritmo implicaran probablemente una redefinición de la base de datos de aprendizaje, incluyendo unos filtros más específicos y estrictos y una definición más detallada de la masa del péptido precursor, que se ha demostrado fundamental para mejorar las predicciones del modelo. Una caracterización de la distribución de los iones separada para diferentes espectrómetros puede mejorar la definición de la función coste, debido al echo que la física subyacente los diferentes proce-

sos de fragmentación y de separación puede producir diferentes patrones espectrales.

La natura de este método, basado en la información global incrustada en el espacio de las configuraciones sintetizada en el perfil de probabilidad de fragmentación, puede ser acoplado a un algoritmo clásico de búsqueda sobre base de datos. Este último usualmente busca el péptido más probable en una base de datos de péptidos según una propia función coste. El potencial *de novo* descrito en este trabajo puede ser aplicado a las secuencias elegidas desde la base de datos como una función coste refinada, creando una nueva puntuación basada en las distribuciones de los señales y del ruido del espectro considerado.

Plegamiento de Proteínas En la segunda parte de esta Tesis hemos aplicado el modelo para el plegamiento de proteínas WSME para describir el comportamiento de la Miotrofina. Esta proteína presenta una estructura en módulos, cada uno de los cuales está formado por dos hélices antiparalelas con cierta estabilidad intrínseca. Los cuatro módulos que componen la molécula se disponen paralelos entre sí formando una estructura lineal. Estas proteínas muestran experimentalmente un plegamiento cooperativo típico de las proteínas globulares que parece en contraste con la modularidad estructural, atrayendo la atención de los investigadores.

Nuestros resultados reproducen cualitativamente el comportamiento experimental de la molécula. El modelo aplicado al análisis del equilibrio y de la cinética del sistema, reproduce la cooperación en ambos ámbitos, y una investigación más profunda muestra que dicha cooperación es compatible con un paisaje de energía libre con múltiples mínimos, predichos por el modelo. La simulación del la cinética del proceso de plegamiento revela la presencia de una heterogeneidad de caminos, como propuesto por Lowe and Itzhaki [42] para explicar el comportamiento de la proteína wild type y de sus mutantes en un único marco. Además, el modelo ha sido aplicado al análisis de algunos mutantes correspondientes a la aplicación de

perturbaciones locales y no locales, permitiendo la predicción cualitativa de comportamientos experimentales (en particular, se muestra que las mutaciones actúan como un interruptor en el flujo hacía un camino u otro de plegamiento y desnaturalización).

Una modificación del modelo para conseguir resultados cuantitativos puede ser introducida ajustando los parámetros del modelo a los datos experimentales disponibles, como mostrado por Bruscolini and Naganathan [7]. Los parámetros son adaptados para reproducir los valores de C_P a través del uso de expresiones fenomenológicas derivadas del mismo modelo WSME. Se ha mostrado que este modelo describe con éxito y cuantitativamente el plegamiento de una proteína de tipo down-hill y una proteína a dos estados[7]. Desafortunadamente, este refinamiento del modelo no puede ser aplicado al caso de la Miotrofina en cuanto se basa en un ajuste a datos calorimétricos que no están disponibles para ella.

En este marco un acercamiento interesante puede ser el estudio de proteínas modulares similares, compuestas por módulos idénticos, las llamadas proteínas modulares de consenso. Estas proteínas están compuestas por repeticiones de una misma cadena de residuos; éstas, aunque non sean proteínas naturales, han sido sintetizadas en laboratorio y sus estructuras han sido resueltas. El estudio de las proteínas de consenso puede aportar a una comprensión mayor de los procesos que gobiernan la dinámica microscópica en los polímeros modulares.

Concluyendo, de acuerdo con las observaciones anteriores, creemos que el uso de simples modelos mecánico-estadísticos nos permitirá producir un cantidad considerable de resultados cualitativos y hasta cuantitativos en problemas de inspiración biológica, en el próximo futuro.

Appendices

Appendix A

Constraints in the state variables

In this appendix we describe the constraints applied to the state-variables $q_\nu^X, \boldsymbol{l}_\nu^X$ and π_ν presented in Section 2.1.

A fragmentation site in ν is characterized by $r_\nu = 0$; the values of the dynamical variables $\sigma_{\nu-1}$ and σ_ν at the sites $\nu - 1$ and ν specify the nature of the residue N-terminal to the fragmentation sites (the "last completed residue"). The fragmentation at ν will produce a number of ions that depend on $\sigma_{\nu-1}$ and σ_ν.

We define the "maximal effective charge", N- or C- terminal to the fragmentation point, q_ν^X $(X = N, C)$, as:

$$q_\nu^X = \min(Q - 1, n_X) \tag{A.1}$$

with Q the total charge of the parent peptide as detected by the first mass spectrometer and n_N, n_C the number of basic residues K, R, H, contained in the N- or C- fragment, respectively. In practice, $q_\nu^X \in [0, Q - 1]$ counts the number of basic residues X-terminal to ν, until this number reaches the

maximal possible value, which is $Q-1$ (since a unit charge is always present from the extra proton attached to the parent peptide) In the fragmentation process, q_ν^X imposes an upper limit on the number of charges species that can be generated: for instance, $q_\nu^N = 0$ and $q_\nu^C = 2$ at some ν for a triply charged spectrum will imply that the N-fragments generated at ν can just have charge +1, while the C-fragments can have charge in the range $[1, 3]$.

The constraints for the values at neighboring sites are as follows:

- If $r_\nu \neq 0$, that is, if ν does not correspond to a peptide bond (a fragmentation sites), or if $r_\nu = 0$ but the residue a ending in ν is not H,K or R, then $q_\nu^X = q_{\nu-1}^X$ (that is $\delta_q^X(a) = 0$) it the only admitted possibility, for both $X = N, C$.

- if $r_\nu = 0$ and the residue a N-terminal to ν is H,K or R, then $q_\nu^N = q_{\nu-1}^N + 1$ (and $\delta_q^N(a) = 1$), provided that $q_{\nu-1}^N < Q - 1$, otherwise, if $q_{\nu-1}^N = Q - 1$, $q_\nu^N = q_{\nu-1}^N$. For the C-term fragments, if $r_\nu = 0$ and the residue a N-terminal to ν is H,K or R, then $q_\nu^C = q_{\nu-1}^C - 1$ ($\delta_q^C(a) = -1$), unless $q_{\nu-1}^C = Q - 1$, in which case also $q_\nu^C = q_{\nu-1}^C$ is allowed, in addition to the already mentioned rule.

The rule on q_ν^C allows to deal with the fact that the number of basic residues C terminal to a point is not known a priori, and not necessarily all of them will be charged.

The constraints on the neutral losses, relating the values of $\boldsymbol{l}_{\nu-1}^X$ and $\boldsymbol{l}_\nu^X$, follow analogous rules of those applied to the charge. If the appended residue a belongs to the set of residues that can loose the neutral group i, then $l_{\nu,i}^N = l_{\nu-1,i}^N + 1$ and $l_{\nu,i}^C = l_{\nu-1,i}^C - 1$, otherwise $l_{\nu,i}^X = l_{\nu-1,i}^X$. Moreover, as in the charge case, a further rule is introduced to respect the maximum expected neutral loss values $\boldsymbol{l}_{\max}$. In this case the latter are parameters of the algorithm, and are not provided together with the parent spectrum. The additional rules affects the case in which $l_{\nu-1,i}^X = l_{\max,i}$. and are completely analogous to the rules for the charge, above.

A further interaction constraint is introduced to reproduce the behaviour of the Trypsin digestion. The outcoming precursor peptides, product of the digestion of the target protein by the Trypsin enzyme, follow a common pattern. Trypsin cleaves the protein at the carboxyl side of the residues Lysine (K) and Arginine (R), while the cleavage is inhibited if the following residue is either a Proline (P) or another K or R. We distinguish, then, between three types of residue: the cleaving residues (K and R), the residues that prevent a previous cleavage (P,K and R), and the rest of residues. The value of π_ν define the nature of the previous residue and force the following residue to respect the tryptic rules: a value $\pi_\nu = 1$ means that the previously added residue was a cleaving one, so that a P,K or R are expected at the next ν where $r_\nu = 0$, unless it is the end of the chain..

The following rules implement the correct constraints:

- if $r_\nu \neq 0$, $\pi_\nu = \pi_{\nu-1}$;
- if $r_\nu = 0$ and a is the species of the residue ending at ν, then
  - if $\pi_{\nu-1} = 0$ and a is not K or R, then $\pi_\nu = 0$,
  - if $\pi_{\nu-1} = 0$ and a is K or R, then $\pi_\nu = 1$,
  - if $\pi_{\nu-1} = 1$ and a is K or R, then $\pi_\nu = 1$,
  - if $\pi_{\nu-1} = 1$ and a is P, then $\pi_\nu = 0$.

Every other combination is forbidden.

The ions generation variables ξ^a_ν only depend on the local state and are non zero if and only if $r_\nu = 0$ and the ion species considered are compatible with the local charge and neutral losses states (i.e. $\xi^{y^{++}}_\nu$ can be 1 only if $q^C_\nu \geq 1$, analogously $\xi^{b-NH_3}_\nu$ can take value 1 if $l^N_{\nu,NH_3} \geq 1$).

In some cases may be useful to keep track of the the number of residues accumulated from the N-term, n_ν. In that case, the constraint is obvious: $n_\nu = n_{\nu-1} + 1$ if $r_n u = 0$, otherwise $n_\nu = n_{\nu-1}$.

Boundaries Conditions. At the N-terminal, at $\nu = 0$, and at the C-terminal, $\nu = M$, the boundaries conditions of the system are characterized by the values of the variables σ_0 and σ_M that correspond to an extremity for the residue sequence. At N-term the first residue starts at $\nu = 0$ with $r_0 = 0$, $n_0 = 1$, and all residue types are admitted and then $\pi_0 = 1, 0$; $q_0^N = 0, \boldsymbol{l}_0^N = \mathbf{0}$, while C-terminal ions can express with the maximum of the charge, $q_0^C = Q - 1$, and, as we do not know the maximum number of neutral losses because we ignore the peptide sequence, then $l_{0,i}^C = 1$ for every i in $[0, l_{\max,i}]$. On the contrary at C-terminal we have $r_M = 0$ but n_M. Analogously to the N-terminal boundaries, here we have $q_M^C = 0$ and $\boldsymbol{l}_M^C = \mathbf{0}$; and for the N-terminal ions we have $q_M^N = Q - 1$ and $l_{M,i}^N = 1$ for $i \in [0, l_{\max,i}]$.

Appendix B

Papers related to this Thesis

- Mauro Faccin, Pierpaolo Bruscolini and Alessandro Pelizzola. Analysis of the Equilibrium and Kinetics of the Ankyrin Repeat Protein Myotrophin. *J. Chem. Phys.*, 134:075102, 2011.
- Mauro Faccin and Pierpaolo Bruscolini. T-novo MS 1: de novo MS/MS spectra interpretation as a statistical mechanical problem. *in preparation*, 2011.
- Mauro Faccin and Pierpaolo Bruscolini. T-novo MS 2: derivation of an effective score function. *in preparation*, 2011.

Bibliography

[1] H. Abe and H. Wako. Analyses of simulations of three-dimensional lattice proteins in comparison with a simplified statistical mechanical model of protein folding. *Phys. Rev. E*, 74(1):011913, Jul 2006. doi: 10.1103/PhysRevE.74.011913.

[2] H. Abe and H. Wako. Folding/unfolding kinetics of lattice proteins studied using a simple statistical mechanical model for protein folding, i: Dependence on native structures and amino acid sequences. *Physica A*, 388(17):3442–3454, SEP 1 2009. ISSN 0378-4371.

[3] Ruedi Aebersold and D.R. Goodlett. Mass spectrometry in proteomics. *Chem. Rev*, 101(2):269–296, 2001.

[4] K. M. Anderson, I. Berrebi-Bertrand, R. B. Kirkpatrick, M. S. McQueney, D. C. Underwood, S. Rouanet, and M. Chabot-Fletcher. cdna sequence and characterization of the gene that encodes human myotrophin/v-1 protein, a mediator of cardiac hypertrophy. *J. Mol. Cell. Cardiol.*, 31(4):705–719, Apr 1999. ISSN 0022-2828. doi: 10.1006/jmcc.1998.0903.

[5] A.E. Ashcroft. *Ionization methods in organic mass spectrometry*. Royal Society of Chemistry, 1997. ISBN 0854045708.

[6] M. Auton, L.M.F. Holthauzen, and D.W. Bolen. Anatomy of energetic

changes accompanying urea-induced protein denaturation. *Proceedings of the National Academy of Sciences*, 104(39):15317, 2007.

[7] Pierpaolo Bruscolini and Athi N Naganathan. Quantitative prediction of protein folding behaviors from a simple statistical model. *J Am Chem Soc*, 133(14):5372–5379, Apr 2011. doi: 10.1021/ja110884m.

[8] Pierpaolo Bruscolini and Alessandro Pelizzola. Exact solution of the muñoz-eaton model for protein folding. *Phys. Rev. Lett.*, 88(25 Pt 1): 258101, Jun 2002. ISSN 0031-9007.

[9] Pierpaolo Bruscolini, Alessandro Pelizzola, and Marco Zamparo. Rate determining factors in protein model structures. *Phys. Rev. Lett.*, 99 (3):038103, Jul 2007. ISSN 0031-9007.

[10] Pierpaolo Bruscolini, Alessandro Pelizzola, and Marco Zamparo. Downhill versus two-state protein folding in a statistical mechanical model. *J. Chem. Phys.*, 126(21):215103, Jun 2007. ISSN 0021-9606. doi: 10.1063/1.2738473.

[11] M. Caraglio, A. Imparato, and Alessandro Pelizzola. Pathways of mechanical unfolding of fniii_{10}: low force intermediates. *J. Chem. Phys.*, 133:065101, 2010.

[12] T. Cellmer, Eric R. Henry, James Hofrichter, and William A. Eaton. Measuring internal friction of an ultrafast-folding protein. *Proc. Natl. Acad. Sci. U. S. A.*, 105(47):18320–18325, NOV 25 2008. ISSN 0027-8424.

[13] Hoi Sung Chung and Andrei Tokmakoff. Temperature-dependent downhill unfolding of ubiquitin. i. nanosecond-to-millisecond resolved nonlinear infrared spectroscopy. *Proteins*, 72(1):474–87, July 2008. ISSN 1097-0134.

[14] Valerie Daggett. Molecular dynamics simulations of the protein unfolding/folding reaction. *Acc Chem Res*, 35(6):422–429, Jun 2002.

[15] V. Dancik, T. Addona, K. Clauser, J. Vath, and Pavel Pevzner. De novo peptide sequencing via tandem mass-spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.

[16] Howard D. Dewald. Electrospray ionization mass spectrometry: Fundamentals, instrumentation and applications (ed. cole, richard b.). *Journal of Chemical Education*, 76(1):33, 1999. doi: 10.1021/ed076p33.1.

[17] J.E. Elias, F.D. Gibbons, O.D. King, F.P. Roth, and S.P. Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, 22(2):214–219, 2004. ISSN 1087-0156.

[18] J.K. Eng, A.L. McCormack, and J.R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, 1994. ISSN 1044-0305.

[19] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, Oct 1989.

[20] Diego U. Ferreiro, Aleksandra M. Walczak, Elizabeth A. Komives, and Peter G. Wolynes. The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLoS Comput. Biol.*, 4(5):e1000070, May 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000070.

[21] R.L. Finley et al. Regulated expression of proteins in yeast using the MAL61-62 promoter and a mating scheme to increase dynamic range. *Gene*, 285(1-2):49–57, 2002. ISSN 0378-1119.

[22] Bernd Fischer, Volker Roth, Franz Roos, Jonas Grossmann, Sacha Baginsky, Peter Widmayer, Wilhelm Gruissem, and Joachim M. Buh-

mann. NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem*, 77(22):7265–7273, 2005.

[23] Ari Frank and Pavel Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem*, 77(4):964–973, 2005.

[24] Y.A. Goo, E.C. Yi, N.S. Baliga, W.A. Tao, M. Pan, Ruedi Aebersold, D.R. Goodlett, L. Hood, and W.V. Ng. Proteomic analysis of an extreme halophilic archaeon, Halobacterium sp. NRC-1. *Molecular & Cellular Proteomics*, 2(8):506, 2003. ISSN 1535-9476.

[25] F. Hillenkamp, M. Karas, R.C. Beavis, and B.T. Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, 63(24):1193–1203, 1991. ISSN 0003-2700.

[26] Franz Hillenkamp, Michael Karas, Ronald C. Beavis, and Brian T. Chait. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Analytical Chemistry*, 63(24):1193A–1203A, 1991. doi: 10.1021/ac00024a002. PMID: 1789447.

[27] D. F. Hochstrasser. Proteome in perspective. *Clin Chem Lab Med*, 36 (11):825–836, Nov 1998. doi: 10.1515/CCLM.1998.146.

[28] A. Imparato and Alessandro Pelizzola. Mechanical unfolding and refolding pathways of ubiquitin. *Phys. Rev. Lett.*, 100(15):158104, Apr 2008. ISSN 0031-9007.

[29] A. Imparato, Alessandro Pelizzola, and Marco Zamparo. Protein mechanical unfolding: a model with binary variables. *J. Chem. Phys.*, 127 (14):145105, Oct 2007. ISSN 0021-9606. doi: 10.1063/1.2776271.

[30] A. Imparato, Alessandro Pelizzola, and Marco Zamparo. Ising-like model for protein mechanical unfolding. *Phys. Rev. Lett.*, 98(14): 148102, Apr 2007.

[31] A. Imparato, Alessandro Pelizzola, and Marco Zamparo. Equilibrium properties and force-driven unfolding pathways of rna molecules. *Phys. Rev. Lett.*, 103(18), OCT 30 2009. ISSN 0031-9007.

[32] Kazuhito Itoh and Masaki Sasai. Cooperativity, connectivity, and folding pathways of multidomain proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 105(37):13865–13870, SEP 16 2008. ISSN 0027-8424.

[33] Kazuhito Itoh and Masaki Sasai. Entropic mechanism of large fluctuation in allosteric transition. *Proc. Natl. Acad. Sci. U. S. A.*, 107(17): 7775–7780, APR 27 2010. ISSN 0027-8424.

[34] R S Johnson and K Biemann. Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed Environ Mass Spectrom*, 18(11):945–57, November 1989. ISSN 0887-6134.

[35] Richard S. Johnson, Stephen A. Martin, Klaus Biemann, John T. Stults, and J. Throck Watson. Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: Differentiation of leucine and isoleucine. *Analitical Chemistry*, 59 (21):2621–2625, Nov 1987. doi: 10.1021/ac00148a019.

[36] Tommi Kajander, Aitziber L. Cortajarena, Ewan R. G. Main, Simon G. J. Mochrie, and Lynne Regan. A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.*, 127(29):10188–10190, Jul 2005. ISSN 0002-7863. doi: 10.1021/ja0524494.

[37] Andrew Keller, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem*, 74 (20):5383–5392, 2002.

[38] Andrew Keller, S. Purvine, Alexey I. Nesvizhskii, S. Stolyar, D.R. Goodlett, and Eugene Kolker. Experimental protein mixture for vali-

dating tandem mass spectral analysis. *OMICS: A Journal of Integrative Biology*, 6(2):207–212, 2002. ISSN 1536-2310.

[39] Sangtae Kim, Nitin Gupta, and Pavel A Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res*, 7(8):3354–3363, Aug 2008. doi: 10.1021/pr8001244.

[40] Ellen Kloss, Naomi Courtemanche, and Doug Barrick. Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Arch. Biochem. Biophys.*, 469(1):83–99, Jan 2008. ISSN 1096-0384. doi: 10.1016/j.abb.2007.08.034.

[41] Alan R. Lowe and Laura S. Itzhaki. Rational redesign of the folding pathway of a modular protein. *Proc. Natl. Acad. Sci. U. S. A.*, 104(8):2679–2684, Feb 2007. ISSN 0027-8424. doi: 10.1073/pnas.0604653104.

[42] Alan R. Lowe and Laura S. Itzhaki. Biophysical characterisation of the small ankyrin repeat protein myotrophin. *J. Mol. Biol.*, 365(4):1245–1255, Jan 2007. ISSN 0022-2836. doi: 10.1016/j.jmb.2006.10.060.

[43] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry*, 17(20):2337–2342, 2003. ISSN 1097-0231.

[44] Michael J. MacCoss, Christine C. Wu, and J.R. Yates III. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem*, 74(21):5593–5599, 2002.

[45] Edward M Marcotte. How do shotgun proteomics algorithms identify proteins? *Nat Biotechnol*, 25(7):755–757, Jul 2007. doi: 10.1038/nbt0707-755.

[46] Edward M. Marcotte, Matteo Pellegrini, Todd O. Yeates, and David Eisenberg. A census of protein repeats. *Journal of Molecular Biology*,

293(1):151 – 160, 1999. ISSN 0022-2836. doi: DOI:10.1006/jmbi.1999.3136.

[47] D.M. Maynard, J. Masuda, X. Yang, J.A. Kowalak, and S.P. Markey. Characterizing complex peptide mixtures using a multi-dimensional liquid chromatography-mass spectrometry system: Saccharomyces cerevisiae as a model system. *Journal of Chromatography B*, 810(1):69–76, 2004. ISSN 1570-0232.

[48] Gerben Menschaert, Tom T M Vandekerckhove, Geert Baggerman, Liliane Schoofs, Walter Luyten, and Wim Van Criekinge. Peptidomics coming of age: a review of contributions from a bioinformatics angle. *J Proteome Res*, 9(5):2051–2061, May 2010. doi: 10.1021/pr900929m.

[49] Philip E. Miller and M. Bonner Denton. The quadrupole mass filter: Basic operating concepts. *Journal of Chemical Education*, 63(7):617–623, July 1986.

[50] Lijuan Mo, Debojyoti Dutta, Yunhu Wan, and Ting Chen. Msnovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal Chem*, 79(13):4870–4878, Jul 2007. doi: 10.1021/ac070039n.

[51] A. N. Morozov, Y. J. Shiu, C. T. Liang, M. Y. Tsai, and S. H. Lin. Nonadditive interactions in protein folding: The zipper model of cytochrome c. *J. Biol. Phys.*, 33(4):255–270, AUG 2007. ISSN 0092-0606.

[52] Leila K. Mosavi, Suzanna Williams, and Zheng yu Peng Zy. Equilibrium folding and stability of myotrophin: a model ankyrin repeat protein. *J. Mol. Biol.*, 320(2):165–170, Jul 2002. ISSN 0022-2836. doi: 10.1016/S0022-2836(02)00441-2.

[53] Leila K. Mosavi, T. J. Cammett, D. C. Desrosiers, and Z. Y. Peng. The ankyrin repeat as molecular architecture for protein recognition.

Protein Sci., 13(6):1435–1448, 2004. ISSN 0961-8368. Times Cited: 211.

[54] D. P. Mukherjee, C. F. McTiernan, and S. Sen. Myotrophin induces early response genes and enhances cardiac gene expression. *Hypertension*, 21(2):142–148, Feb 1993.

[55] Victor Muñoz and William A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U. S. A.*, 96(20):11311–11316, September 1999. ISSN 0027-8424.

[56] Victor Muñoz, P A Thompson, James Hofrichter, and William A. Eaton. Folding dynamics and mechanism of beta-hairpin formation. *Nature*, 390(6656):196–199, November 1997. ISSN 0028-0836.

[57] Victor Muñoz, Eric R. Henry, James Hofrichter, and William A. Eaton. A statistical mechanical model for beta-hairpin kinetics. *Proc. Natl. Acad. Sci. U. S. A.*, 95(11):5872–5879, May 1998. ISSN 0027-8424.

[58] C. Narasimhan, D.L. Tabb, N.C. VerBerkmoes, M.R. Thompson, R.L. Hettich, and E.C. Uberbacher. MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Analytical chemistry*, 77(23):7581–7593, 2005. ISSN 0003-2700.

[59] Alexey I. Nesvizhskii, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry*, 75(17):4646–4658, 2003. ISSN 0003-2700.

[60] Adrian A Nickson and Jane Clarke. What lessons can be learned from studying the folding of homologous proteins? *Methods*, 52(1):38–50, Sep 2010. doi: 10.1016/j.ymeth.2010.06.003.

[61] Alessandro Pelizzola. Exactness of the cluster variation method and factorization of the equilibrium probability for the wako-saito-munoz-eaton model of protein folding. *J. Stat. Mech.-Theory Exp.*, 11010: 11010, 2005.

[62] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, Dec 1999. doi: 3.0.CO;2-2.

[63] John T Prince, Mark W Carlson, Rong Wang, Peng Lu, and Edward M Marcotte. The need for a public proteomics repository. *Nat Biotech*, 22(4):471–472, April 2004. ISSN 1087-0156.

[64] P Roepstorff and J Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom*, 11 (11):601, nov 1984. ISSN 0306-042X.

[65] L.M. Schnapp, S. Donohoe, J. Chen, D.A. Sunde, P.M. Kelly, J. Ruzinski, T. Martin, and D.R. Goodlett. Mining the acute respiratory distress syndrome proteome: identification of the insulin-like growth factor (IGF)/IGF-binding protein-3 pathway in acute lung injury. *American Journal of Pathology*, 169(1):86, 2006.

[66] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978. ISSN 0090-5364.

[67] S. Sen, G. Kundu, N. Mekhail, J. Castel, K. Misono, and B. Healy. Myotrophin: purification of a novel peptide from spontaneously hypertensive rat heart that influences myocardial growth. *J. Biol. Chem.*, 265(27):16635–16643, Sep 1990. ISSN 0021-9258.

[68] C.E. Shannon. The mathematical theory of communication (parts 1 and 2). *Bell Syst. Tech. J*, 27:379–423, 1948.

[69] N. Sivasubramanian, G. Adhikary, P. C. Sil, and S. Sen. Cardiac myotrophin exhibits rel/nf-kappa b interacting activity in vitro. *J. Biol. Chem.*, 271(5):2812–2816, Feb 1996. ISSN 0021-9258.

[70] Zoltán Takáts, Justin M Wiseman, Bogdan Gologan, and R. Graham Cooks. Electrosonic spray ionization. a gentle technique for generating folded proteins and protein complexes in the gas phase and for studying ion-molecule reactions at atmospheric pressure. *Anal Chem*, 76(14): 4050–4058, Jul 2004. doi: 10.1021/ac049848m.

[71] M. Taoka, T. Isobe, T. Okuyama, M. Watanabe, H. Kondo, Y. Yamakawa, F. Ozawa, F. Hishinuma, M. Kubota, and A. Minegishi. Murine cerebellar neurons express a novel gene encoding a protein related to cell cycle control and cell fate determination proteins. *J. Biol. Chem.*, 269(13):9946–9951, Apr 1994. ISSN 0021-9258.

[72] J.A. Taylor and R.S. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9):1067–1075, 1997. ISSN 1097-0231.

[73] J.A. Taylor and R.S. Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem*, 73(11):2594–2604, 2001.

[74] Yuzo Ueda, Hiroshi Taketomi, and Nobuhiro Gō. Studies on protein folding, unfolding, and fluctuations by computer simulation. ii. a. three-dimensional lattice model of lysozyme. *Biopolymers*, 17(6):1531–1548, 1978. ISSN 1097-0282.

[75] H. Wako and N. Saito. Statistical mechanical theory of protein conformation .1. general considerations and application to homopolymers. *J. Phys. Soc. Japan*, 44(6):1931–1938, 1978.

[76] H. Wako and N. Saito. Statistical mechanical theory of protein con-

formation .2. folding pathway for protein. *J. Phys. Soc. Japan*, 44(6): 1939–1945, 1978.

[77] Yunhu Wan, Austin Yang, and Ting Chen. Pephmm: a hidden markov model based scoring function for mass spectrometry database search. *Anal Chem*, 78(2):432–437, Jan 2006. doi: 10.1021/ac051319a.

[78] Svava K. Wetzel, Giovanni Settanni, Manca Kenig, H. Kaspar Binz, and Andreas Plückthun. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J. Mol. Biol.*, 376(1): 241–257, Feb 2008. ISSN 1089-8638. doi: 10.1016/j.jmb.2007.11.046.

[79] M. R. Wilkins, J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser, and K. L. Williams. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev*, 13:19–50, 1996.

[80] M. Yamashita and JB Fenn. Electrospray 11. ion source: another variation of the freejet theme. *Phys. Chem*, 88:4451–4459, 1984.

[81] Y. Yang, S. Nanduri, S. Sen, and J. Qin. The structural basis of ankyrin-like repeat function as revealed by the solution structure of myotrophin. *Structure*, 6(5):619–626, May 1998.

[82] J.R. Yates III, J.K. Eng, and A.L. McCormack. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Analytical Chemistry*, 67(18):3202–3210, 1995. ISSN 0003-2700.

[83] J.R. Yates III, J.K. Eng, A.L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry*, 67(8):1426–1436, 1995. ISSN 0003-2700.

[84] Marco Zamparo and Alessandro Pelizzola. Rigorous results on the local equilibrium kinetics of a protein folding model. *J. Stat. Mech.-Theory Exp.*, page P12009, 2006. ISSN 1742-5468.

[85] Marco Zamparo and Alessandro Pelizzola. Kinetics of the wako-saitô-muñoz-eaton model of protein folding. *Phys. Rev. Lett.*, 97(6):068106, Aug 2006. ISSN 0031-9007.

[86] Marco Zamparo and Alessandro Pelizzola. Nearly symmetrical proteins: folding pathways and transition states. *J. Chem. Phys.*, 131(3): 035101, Jul 2009. ISSN 1089-7690. doi: 10.1063/1.3170984.

[87] Zhongqi Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal Chem*, 76(14):3908–3922, Jul 2004. doi: 10.1021/ac049951b.