# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

186,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities


WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

## Interested in publishing with us?
## Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

7

# Computational Methods in Mass Spectrometry-Based Protein 3D Studies

Rosa M. Vitale[1], Giovanni Renzone[2], Andrea Scaloni[2] and Pietro Amodeo[1]
[1]*Istituto di Chimica Biomolecolare, CNR, Pozzuoli*
[2]*Laboratorio di Proteomica e Spettrometria di Massa, ISPAAM, CNR, Naples*
*Italy*

## 1. Introduction

Mass Spectrometry (MS)-based strategies featuring chemical or biochemical probing represent powerful and versatile tools for studying structural and dynamic features of proteins and their complexes. In fact, they can be used both as an alternative for systems intractable by other established high-resolution techniques, and as a complementary approach to these latter, providing different information on poorly characterized or very critical regions of the systems under investigation (Russell et al., 2004). The versatility of these MS-based methods depends on the wide range of usable probing techniques and reagents, which makes them suitable for virtually any class of biomolecules and complexes (Aebersold et al., 2003). Furthermore, versatility is still increased by the possibility of operating at very different levels of accuracy, ranging from qualitative high-throughput fold recognition or complex identification (Young et al., 2000), to the fine detail of structural rearrangements in biomolecules after environmental changes, point mutations or complex formations (Nikolova et al.,1998; Millevoi et al., 2001; Zheng et al., 2007). However, these techniques heavily rely upon the availability of powerful computational approaches to achieve a full exploitation of the information content associated with the experimental data.
The determination of three-dimensional (3D) structures or models by MS-based techniques (MS3D) involves four main activity areas: 1) preparation of the sample and its derivatives labelled with chemical probes; 2) generation of derivatives/fragments of these molecules for further MS analysis; 3) interpretation of MS data to identify those residues that have reacted with probes; 4) derivation of 3D structures consistent with information from previous steps. Ideally, this procedure should be considered the core of an iterative process, where the final model possibly prompts for new validating experiments or helps the assignment of ambiguous information from the mass spectra interpretation step.
Both the overall MS3D procedure and its different steps have been the subject of several accurate review and perspective articles (Sinz, 2006; Back et al., 2003; Young et al., 2000; Friedhoff, 2005, Renzone, et al., 2007a). However, with the partial exception of a few recent papers (Van Dijk et al., 2005; Fabris et al., 2010; Leitner et al., 2010), the full computational detail behind 3D model building (step 4) has generally received less attention than the former three steps. Structural derivation in MS3D, in fact, is considered a special case of structural determination from sparse/indirect constraints (SD-SIC). Nevertheless, information for modelling derivable from MS-based experiments exhibits some peculiar

features that differentiate it from the data types associated with other experimental techniques involved in SD-SIC procedures, such as nuclear magnetic resonance (NMR), electron microscopy, small-angle X-ray scattering (SAXS), Förster resonance energy transfer (FRET) and other fluorescence spectroscopy techniques, for which most of the currently available SD-SIC methods have been developed and tailored (Förster et al., 2008; Lin et al., 2008; Nilges et al., 1988a; Aszodi et al., 1995).

In this view, this study will illustrate possible approaches to model building in MS3D, underlining the main issues related to this specific field and outlining some of the possible solutions to these problems. Whenever possible, alternative methods employing either different programs selected among most popular applications in homology modelling, threading, docking and molecular dynamics (MD), or different strategies to exploit the information contained in MS data will be described. Discussion will be limited to packages either freely available, or costing less than 1,000 US$ for academic users. For programs, the home web address has been reported, rather than references that are very often partial and/or outdated. Some examples, derived from the literature available in this field, or developed *ad hoc* to illustrate some critical features of the computational methods in MS3D, should clarify potentiality and current limitations of this approach.

## 2. General MS3D modelling procedures

### 2.1 Possible computational protocols for MS3D approaches

MS3D can be fruitfully applied to many structure-related problems; thus, it requires the (possibly combined) use of different modelling procedures. However, a very general scheme for a MS3D approach can still be sketched (Fig. 1). It includes:

- an initial generation of possible structures for the investigated system by some sampling algorithms (S1 or S2 stages);
- followed by classification, clustering and selection steps of the best sampled structures based on one or more criteria (F1 or F2a-F2b-F2c);
- an optional narrowing of the ensemble by a refinement of the selected models (R);
- followed by new classification, clustering and selection stages for the identification of the most representative models (FF).

Selection criteria are very often represented by more or less sophisticated combinations of different scoring (i.e. the higher, the better), penalty (i.e. the lower, the better) or target (i.e. the closer to its reference value, the better) functions. For the sake of brevity, from here onwards the term "scoring" will be indiscriminately used for either true scoring, or penalty, or target function, when their discrimination is not necessary.

The features characterizing a specific approach are: a) combination of sampling (and optimization) algorithms, b) scoring functions in sampling/optimization and classification/clustering/selection stages, c) strategies to introduce MS-based experimental information.

A first major branching in this scheme already occurs in the earliest modelling stages (box A), depending if MS-based information is, at least in part, integrated in the structure generation stage (path S1-F1), or rather deferred to a subsequent model classification/selection step (path S2-F2a-F2b-F2c).

Depending on information types, programs and strategies used in modelling (see next sections for theory and examples), MS-based data can be either all introduced during sampling (S1), or all used in the filtering stage (F2a), or subdivided between the two steps (S1+F1). The main advantage of the inclusion of MS-based information into sampling (path

S1-F1) is an increase in model generation efficiency by limitation of the conformational or configurational subspace to be explored. In several potentially problematic cases, i.e. large molecules with very limited additional information available, this reduction can transform a potentially insoluble problem into a reliable model generation, capable of correlating structural and functional features of the investigated system. However, for the very same reason, if information is introduced too abruptly or tightly during structural sampling, it can artificially freeze the models into a wrong, or at least incomplete, set of solutions (Latek et al., 2007; Bowers et al., 2000). Also the weight of erroneous restraints will be considerably amplified by the impossibility of a comparison with solutions characterized by some restraint violations, but considerably more favourable scoring function values, which are often diagnostic of inadequate sampling and/or errors in the experimental restraint set.

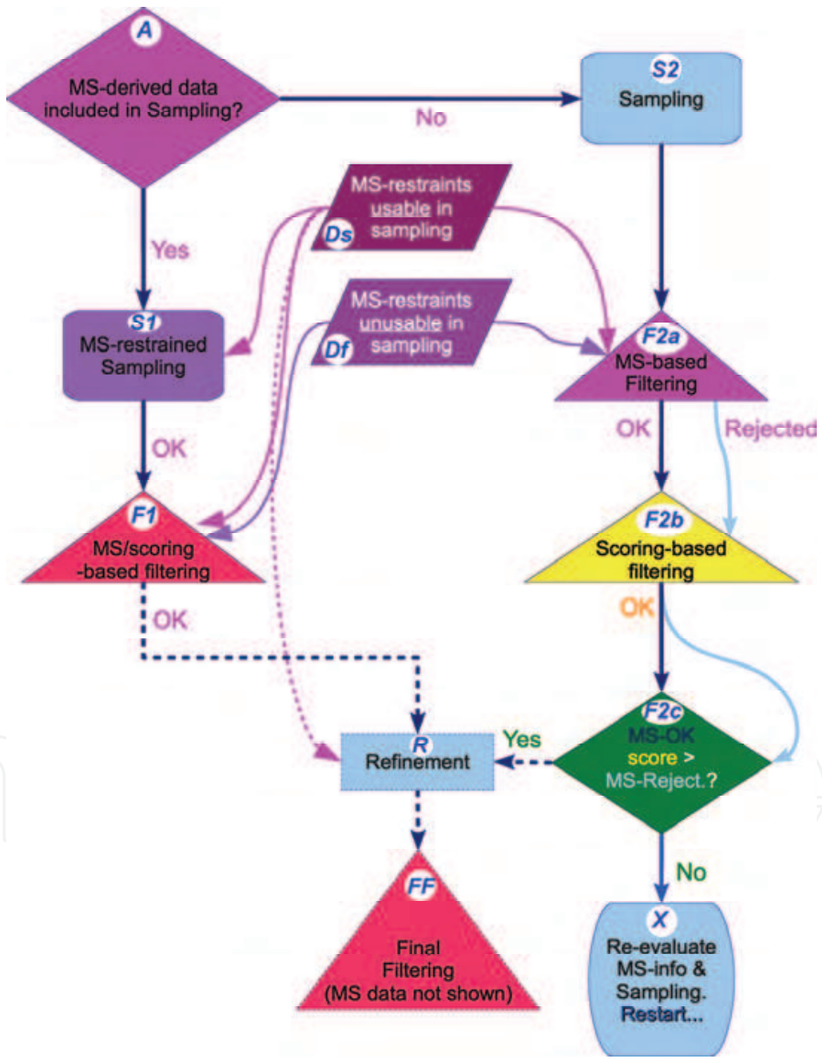



Fig. 1. Flowchart of a generic MS3D modelling approach. Magenta, violet and pink represent steps in which MS-based information is applied. Triangular arrows indicate use of MS-based data. Dotted lines and borders are used for optional refinement stages. Blue codes in white circles/ellipses label the corresponding stages within the text.

Accordingly, both the protocol used to implement MS-based information into modelling procedures and the MS-based data themselves generally represent very critical features, which require the maximum attention during computational setup and final analyses. In addition, implementation of restraints in the sampling procedure either requires some purposely programming activity, or severely limits the choice of modelling tools to programs already including suitable user-defined restraints.
Use of MS-based information in post-sampling analyses (path S2-F2a-F2b-F2c) to help classifying and selecting the final models exhibits a mostly complementary profile of advantages-disadvantages. In fact, it decreases the sampling efficiency of the modelling methods (S2), by leading to a potentially very large number of models to be subsequently discarded on the mere basis of their violations of MS-derived restraints (F2a), and by providing no *ab initio* limitations to the available conformational/configurational space of the system. Furthermore, it may still require programming activity if available restraint analysis tools (F2a) are lacking or inefficient in the case of the implemented information. However, this approach warrants the maximum freedom to the user in the choice of the sampling program; this may result very useful in those cases where the peculiar features of a specific program are strongly required to model the investigated system. In addition, a compared analysis of both structural features and scoring function values between models accepted and rejected on the basis of MS-based data may allow the identification of potential issues in the selected models and the corresponding data sets (steps F2c-X).

### 2.2 Integration of MS-based data into modelling procedures

Although an ever-increasing number of MS-based strategies has been developed, they provide essentially two information classes for model building: i) surface accessible residues, from chemical/isotopic labelling or limited proteolysis experiments (Renzone et al., 2007a); ii) couples of residues whose relative distances span a prefixed range, from crosslinking experiments (Sinz, 2006; Renzone et al., 2007a). Details on the nature of the combined biochemical and MS approaches used to generate these data and the experimental procedures adopted in these cases is provided in the exhaustive reviews reported above.

#### 2.2.1 Surface-related information (selective proteolysis and chemical labelling)

Although many structural generation approaches include surface-dependant terms, usually they are not exposed to the user; thus, direct implementation of accessibility information is always indirect and ranges from very difficult to impossible. In some docking programs, surface residue patches can be excluded from the exploration, thus restricting the region of space to be sampled (Section 3.2). This information is generally exploited through programs that build and evaluate different kinds of molecular surfaces, applied during the model validation stages. In this view, the main available programs and their usage will be described in the section dedicated to model validation (Section 3.3.2).
In the case of modelling procedures based on sequence alignment with templates of known 3D structure, surface-dependent data can be employed both to validate alignments before modelling (early steps in S1 stage), and to filter the structures resulting from the different steps of a traditional model building procedure (stages F1 or F2a, and FF).

#### 2.2.2 Crosslinks

Cross-linking information often directly contribute to the model building procedure (under the form of distance restraints or direct linker addition to the simulated systems) (stage S1 in Fig.1), in addition to their model validation/interpretation role (stages F1, F2a, FF).

Whenever information from crosslinking experiments is integrated within the modelling procedure, the most common approach recurring in literature is its translation into distance constraints (i.e. "hard", fixed distances) or restraints (i.e. variable within an interval and/or around a fixed distance with a given tolerance) involving atoms, in a full-atomistic representation, or higher-order units, such as residues, secondary structure (SS) elements, or domains, in coarse-grained models. A less common approach consists in the explicit inclusion of the crosslinker atoms in the simulation.

#### 2.2.2.1 Distance restraints

Distance restraints (DRs) are usually implemented by adding a penalty term to the scoring function used to generate, classify or select the models, whenever the distance between specified atom pairs exceeds a threshold value. In this way, associated experimental information can be introduced rather easily and with moderate computational overheads in all the molecular modelling and simulation approaches based on scoring functions. However, since crosslinking agents are molecules endowed with well-defined and specific conformational and interaction properties, both internal and with crosslinked molecules, accurate theoretical and experimental estimates of distance ranges associated with the corresponding cross-link agents only qualitatively correspond to experimentally-detected distances between pairs of cross-linked residues (Green et al., 2001; Leitner et al., 2010). Steric bumps, specific favourable or unfavourable electrostatic interactions, presence of functional groups capable of promoting/hampering the crosslinking reaction and changes in crosslinker conformational population under the effects of macromolecule are all possible causes for observed discrepancies.

#### 2.2.2.2 Explicit linkers

Explicit inclusion of crosslinkers in the systems, although potentially allowing to overcome the limits of DRs, presently suffers from several drawbacks that limit its usage to either final selection/validation stages, or to cases where a limited number of totally independent and simultaneously holding crosslinks are observed. In fact, when many crosslinks are detected in a system by MS analysis, they very often correspond to mixtures of different patterns, because crosslinks can interfere each other either by direct steric hindrance, or by competition for one of the macromolecule reacting groups, or by inducing deformation in the linked system, thus preventing further reactions. However, the added information from explicit crosslinkers may: i) allow disambiguation between alternative predicted binding modes, ii) provide more realistic and strict estimates of the linker length to be used in further stages of DR-based calculations, iii) help modelling convergence, iv) substantially contribute to model validation.
An attempt to reproduce by an implicit approach at least the geometrical constraints associated with a physical linker has been performed by developing algorithms to identify minimum-length paths on protein surfaces (Potluri et al., 2004). This approach provides upper/lower bounds to possible crosslinking distances on static structures but it only worked on static structures as a post-modelling validation tool, and no further applications have been reported so far.

## 3. Available computational approaches in MS3D

MS-based data can be used to obtain structural information on different classes of problems:
a. single conformational states (e.g. the overall fold);

b. conformational changes upon mutations/environmental modifications;
c. macromolecular aggregation (multimerization);
d. binding of small ligands to macromolecules.

Sampling efficiency and physical soundness of the scoring functions used during sampling (stages S1/S2 of Fig. 1) and to select computed structures (stages F1/F2b and FF) generally represent the main current limitations of 3D structure prediction and simulation methods. In this view, introduction of experimental data represents a powerful approach to reduce the geometrical space to be explored during sampling, and also an independent criterion to evaluate the quality of selected models.

From a computational point of view, structural problems a)-d) translate into system-dependent proper combinations of:

A. fold identification and characterization;
B. docking;
C. structural refinement and characterization of dynamic properties and of changes under the effects of local or environmental perturbations.

Since the optimal combination of methods for a given problem depends upon a large number of system- and data-dependent parameters, and the number of programs developed for biomolecular simulations is huge, an exhaustive description and compared analysis of methods for biomolecular structure generation/refinement is practically impossible. However, we will try to offer a general overview of the main approaches to generate, refine and select 3D structures in MS3D applications, with a special attention to possible ways of introducing MS-based data and exploiting their full information content.

### 3.1 Fold identification and characterization

The last CASP (Critical Assessment of techniques for protein Structure Prediction) experiment call (CASP9, 2010) classified modelling methods in two main categories: "Template Based Modelling" (TBM) and "Template Free Modelling" (TFM), depending if meaningful homology can be identified or not before modelling between the target sequence and those of proteins/domains whose 3D structures are known (templates).

TFM represents the most challenging task because it requires the exploration of the widest conformational space and heavily relies on scoring methods inspired by those principles of physics governing protein folding (*de novo* or *ab initio* methods), eventually integrated by statistical predictions, such as probabilities of interresidue contacts, surface accessibility of single residues or local patches and SS occurrence. When number and quality of these information increase, together with the extent of target sequence for which they are available, "folding recognition" and "threading" techniques can be used, including a broad range of methods at the interface between TFM and TBM. In these approaches, several partial 3D structure "seeds" are generated by statistical prediction or distant homology relationships, and their relative arrangements are subsequently optimized by strategies deriving from *de novo* methods.

The most typical TBM approach, "comparative" or "homology" modelling (HM), uses experimentally elucidated structures of related protein family members as "templates" to model the structure of the protein under investigation (the "target"). Target sequence can either be fully covered by one or more templates, exhibiting good homology over most of the target sequence, or can require a "patchwork" of different templates, each best covering a different region of the target.

A further group of approaches, presently under active development and already exhibiting good performances in CASP and other benchmark and testing experiments, is formed by the "integrative" or "hybrid" methods. They combine information from a varied set of computational and experimental sources, often acting as/based on "metaservers", i.e. servers that submit a prediction request to several other servers, then averaging their results to provide a consensus that in many cases is more reliable than the single predictions from which it originated. Some metaservers use the consensus as input to their own prediction algorithms to further elaborate the models.

In order to provide some guidelines for structural prediction/refinement tasks in the presence of MS-based data, a general procedure will be outlined for protein fold/structure modelling. The starting step in protein modelling is usually represented by a search for already structurally-characterized similar sequences. Sensitive methods for sequence homology detection and alignment have been developed, based on iterative profile searches, e.g. PSI-Blast (Altschul et al., 1997), Hidden Markov Models, e.g. SAM (K. Karplus et al. 1998), HMMER (Eddy, 1998), or profile-profile alignment such as FFAS03 (Jaroszewski et al., 2005), profile.scan (Marti-Renom et al., 2004), and HHsearch (Soding, 2005).

When homology with known templates is over 40%, HM programs can be used rather confidently. In this case, especially when alignments to be used in modelling have already been obtained, local programs represent a more viable alternative to web-based methods than in TFM processes. If analysis is limited to most popular programs and web services capable of implementing user MS-based restraints (strategy S1 in Fig. 1), the number of possible candidates considerably decreases. Among web servers, on the basis of identified homologies with templates, Robetta is automatically capable of switching from *ab initio* to comparative modelling, while I-TASSER requires user-provided alignment or templates to activate comparative modelling mode. A very powerful, versatile and popular HM program, available both as a standalone application, and as a web service, and embedded in many modelling servers, is MODELLER (http://www.salilab.org/modeller/). It include routines for template search, sequence and structural alignments, determination of homology-derived restraints, model building, loop modelling, model refinement and validation. MS-based distance restraints can be added to those produced from target-template alignments, as well as to other restraints enforcing secondary structures, symmetry or part of the structure that must not be allowed to change upon modelling. However, some scripting ability is required to fully exploit MODELLER versatility.

The overall accuracy of HM models calculated from alignments with sequence identities of 40% or higher is almost always good (typical root mean square deviations (RMSDs) from corresponding experimental structures less than 2Å). The frequency of models deviating by more than 2Å RMSD from experimental structures rapidly increases when target–template sequence identity falls significantly below 30–40%, the so-called "twilight zone" of HM (Blake & Cohen, 2001; Melo & Sali, 2007). In such cases, the quality of resulting modelled structures significantly increases by combining additional information, both of statistical origin, such as SS prediction profiles, and from sparse experimental data (low resolution NMR or chemical crosslinking, limited proteolysis, chemical/isotopical labelling coupled with MS).

If the search does not produce templates with sufficient homology and/or covering of the target sequence, TFM or mixed TFM/TBM methods must be used. Many programs based on *ab initio*, fold recognition and threading methods are presently offered as web services; this is because very often they use a metaserver approach for some steps, need extensive

searches in large databases, require huge computational resources, or to better protect underlying programs and algorithms, currently under very active development. Although this may offer some advantages, especially to users less-experienced in biocomputing or endowed with limited computing facilities, it may also imply strong limitations in the full exploitation of the features implemented in the different methods, with particularly serious implications in MS3D. Only few servers either include a NMR structure determination module (not always suitable for MS-based data), or explicitly allow the optional usage of user-provided distance restraints in the main input form. Fortunately, two of the most used and versatile servers, Robetta (http://robetta.bakerlab.org/) and I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/), good performers at the last CASP rounds (http://predictioncenter.org/), allow input of distance restraints in the modelling procedure, via a NMR-dedicated service for Robetta (Rosetta-NMR, suitable for working with sparse restraint) (Bowers et al., 2000), or directly in the main prediction submission page (I-TASSER). Other servers can still allow the implementation of MS-based information in the model generation step if they can save intermediate results, such as sequence alignments, SS or fold predictions. These latter, after addition of MS-based restraints, can be then included into suitable modelling programs, to be run either locally or on web servers.

A successful examples of modelling with MS-based information in a low-homology case is Gadd45β. A model was built, despite the low sequence identity (<20%) with template identified by fold recognition programs, through the introduction of additional SS restraints, which were based on SS profiles and experimental data from limited proteolysis and alkylation reactions combined with MS analysis (Papa et al., 2007). Model robustness was confirmed by comparison with the homolog Gadd45γ structure solved later (Schrag JD et al., 2008), where the only divergence in SS profiles was the occurrence of two short $3_{10}$ helices (three residues each long) and an additional two-residues β-strand in predicted loop regions (Fig. 2). Furthermore, this latter β-strand is so distorted that only a few SS assignment programs could identify it, and the corresponding sequence in Gadd45β, predicted unstructured and outside the template alignment, was not modelled at all.

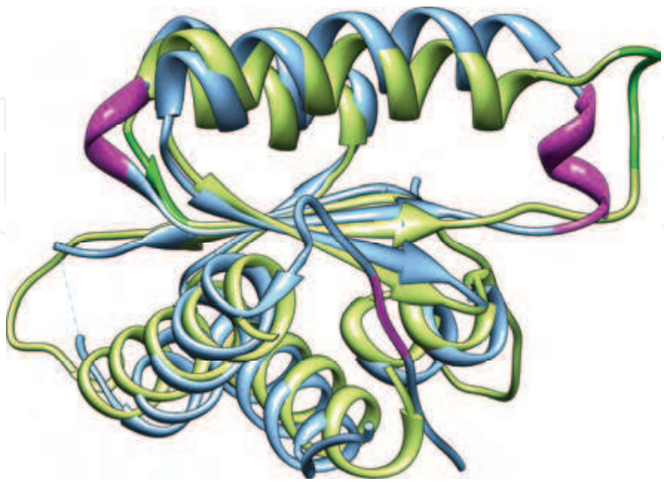


Fig. 2. Comparison between the MS3D model of Gadd45β (light green) and the crystallographic structure of its homolog Gadd45γ (light blue). Sequences with different SS profiles are painted green in Gadd45β and magenta in Gadd45γ.

### 3.2 Docking

Usually, methods for protein docking involve a six-dimensional search of the rotational and translational space of one protein with respect to the other where the molecules are treated as rigid or semirigid-bodies. However, during protein-protein association, the interface residues of both molecules may undergo conformational changes that sometimes involve not only side-chains, but also large backbone rearrangements. To manage at least in part these conformational changes, protein docking protocols have introduced some degree of protein flexibility by either use of "soft" scoring functions allowing some steric clash, or explicit inclusion of domain movement/side chain flexibility. Biological information from experimental data on regions or residues involved in complexation can guide the search of complex configurations or filter out wrong solutions. Among the programs most frequently used for protein-protein docking, recently reviewed by Moreira and colleagues (Moreira et al., 2010), some of them can manage biological information and will be discussed in this context.

In the Attract program (http://www.t38.physik.tu-muenchen.de/08475.htm ), proteins are represented with a reduced model (up to 3 pseudoatoms per amino acid) to allow the systematic docking minimization of many thousand starting structures. During the docking, both partner proteins are treated as rigid-body and the protocol is based on energy minimization in translational and rotational degrees of freedom of one protein with respect to the other. Flexibility of critical surface side-chains as well as large loop movements are introduced in the calculation by using a multiple conformational copy approach (Bastard et al., 2006). Experimental data can be taken into account at various stages of the docking procedure.

The 3D-Dock algorithm (http://www.sbg.bio.ic.ac.uk/docking/) performs a global scan of translational and rotational space of the two interacting proteins, with a scoring function based on shape complementarity and electrostatic interaction. The protein is described at atomic level, while the side-chain conformations are modelled by multiple copy representation using a rotamer library. Biological information can be used as distance restraints to filter final complexes.

HADDOCK (http://www.nmr.chem.uu.nl/haddock/) makes use of biochemical or biophysical interaction data, introduced as ambiguous intermolecular distance restraints between all residues potentially involved in the interaction. Docking protocol consists of four steps: 1) topology and structure generation; 2) randomization of orientations and rigid body energy minimization; 3) semi-flexible simulated annealing (SA) in torsion angle space; 4) flexible refinement in Cartesian space with explicit solvent (water or DMSO). The final structures are clustered using interface backbone RMSD and scored by their average interaction energy and buried interface area. Recently, also explicit inclusion of water molecules at the interface was incorporated in the protocol.

Molfit (http://www.weizmann.ac.il/Chemical_Services/molfit/) represents each molecule involved in docking process by a 3-dimensional grid of complex numbers and estimates the extent of geometric and chemical surface complementarity by correlating the grids using Fast Fourier Transforms (FFT). During the search, contacts involving specified surface regions of either one or both molecules are up- or down-weighted, depending on available structural and biochemical data or sequence analysis (Ben-Zeev et al., 2003). The solutions are sorted by their complementarity scores and the top ranking solutions are further refined by small rigid body rotations around the starting position.

PatchDock (http://bioinfo3d.cs.tau.ac.il/PatchDock/) is based on shape complementarity. First, the surfaces of interacting molecules are divided according to the shape in concave, convex and flat patches; then, complementarity among patches are identified by shape-matching techniques. The algorithm is a rigid body docking, but some flexibility is indirectly considered by allowing some steric clashes. The resulting complexes are ranked on the basis of the shape complementarity score. PatchDock allows integration of external information by a list of binding site residues, thus restricting the matching stage to their corresponding patches.

RosettaDock (http://rosettadock.graylab.jhu.edu/) try to mimics the two stages of a docking process, recognition and binding, as hypothesized in Camacho & Vajda, 2001. Recognition is simulated by a low resolution phase in which a coarse-grained representation of proteins, with side chains replaced by single pseudoatoms, undergoes a rigid body Monte Carlo (MC) search on translations and rotations. Binding is emulated by a high-resolution refinement phase where explicit sidechains are added by using a backbone-dependent rotamer packing algorithm. The sampling problem is handled by supercomputing clusters to ensure a very large number of decoys that are discriminated by scoring functions at the end of both stages of docking. The docking search problem can be simplified when biological information is available on the binding region of one or both interacting proteins. The reduction of conformational space to be sampled could be pursued by: i) opportunely pre-orienting the partner proteins, or ii) reducing docking sampling to the high-affinity domain, in the case of multidomain proteins, or iii) using loose distance constraints.

ZDOCK (http://zdock.bu.edu/) is a rigid body docking program based on FFT algorithm and an energy function that combines shape complementarity, electrostatics and desolvation terms. RDOCK (http://zdock.bu.edu/) is a refinement program to minimize and rerank the solutions found by ZDOCK. The complexes are minimized by CHARMm (Brooks et al., 1983) to remove clashes and improve energies, then electrostatic and desolvation terms are recalculated in a more accurate fashion with respect to ZDOCK. Biological information can be used either to avoid undesirable contacts between certain residues during ZDOCK calculations or to filter solutions after RDOCK.

As in protein folding, also for docking the use of MS-based information allowed the modelling of several complexes even in the lack of suitable templates with high homology. The fold of prohibitin proteins PHB1 and PHB2 was predicted (Back et al., 2002) by SS and fold recognition algorithms, while crosslinking allowed to model the relative spatial arrangement of the two proteins in their 1:1 complex. Another example of combined use of SS information, chemical crosslinking, limited proteolysis and MS analysis results with a low sequence identity (~ 20%) template is the modelling of porcine aminoacylase 1 dimer; in this case, standard modelling procedures based on automatic alignment had failed to produce a dimeric model consistent with experimental data (D'Ambrosio et al., 2003).

In the case of protein-small ligand docking, the conformational space to be explored is reduced by the small size of the ligand, whose full flexibility can usually be allowed, and by the limited fraction of protein surface to be sampled, corresponding to the binding site, often already known. Among the programs for ligand-flexible docking that allow protein side-chains flexibility, Autodock is one of most popular (http://autodock.scripps.edu/). AutoDock combines a grid-based method with a Lamarckian Genetic Algorithm to allow a rapid evaluation of the binding energy. A simulated annealing method and a traditional genetic algorithm are also available in Autodock4.

In general, MS-based data can be used to limit the protein region to be sampled (Kessl et al., 2009) or can be explicitly considered in the docking procedure, as in the case of the mapping of Sso7d ATPase site (Renzone et al., 2007b). In this case, three independent approaches for molecular docking/MD studies were followed, considering both FSBA-derivatives and the ATP-Sso7d non-covalent complex: i) unrestrained MD, starting from a full-extended, external conformation for Y7-FSBA and K39-FSBA residue sidechains, and from several random orientations for ATP, with an initial distance of 20 Å from Sso7d surface, in regions not involved in protein binding; ii) restrained MD, by gradually imposing distance restraints corresponding to a H-bond between adenine $NH_2$ group and each accessible (i.e., within a distance lower or equal to the maximum length of the corresponding FSBA-derivative) donor sidechain; iii) rigid ligand docking, by calculating 2000 ZDOCK models of the non-covalent complex of Sso7d with an adenosine molecule. The rigid ligand docking reproduced only in part features from other approaches, as rigid docking correctly predicted the anchoring point for adenosine ring, but failed to achieve a correct position for the ribose moiety, due to the required concerted rearrangement of two Sso7d loops involved in the binding. This latter feature represents one of the main advantages of modelling strategies involving MD (in particular, in cartesian coordinates) because MD-based simulation techniques are the best or the only approaches that reproduce medium-to-large scale concerted rearrangements of non-contiguous regions.

### 3.3 Model simulation, refinement and validation

Refinement (R stage in Fig.1) and validation of final models (FF stage) represent very important steps, especially in cases of low homologies with known templates and when fine details of the models are used to predict or explain functional properties of the investigated system. In addition, very often the modelled structures are aimed at understanding the structural effects of point mutations or other local sequence alterations (sequence deletions/insertions, addition or deletion of disulphide bridges, formation of covalent constructs between two molecules and post-translational modifications), or of changes in environmental parameters (temperature, pressure, salt concentration and pH). In these cases, techniques are required to simulate the static or dynamic behaviour of the investigated system in its perturbed and unperturbed states.

#### 3.3.1 Computational techniques and programs for model simulation and refinement

Model refinement, when not implemented in the modelling procedure, can be performed by energy minimization (EM) or, better, by different molecular simulation methods, mostly based on variants of molecular dynamics (MD) or Monte Carlo (MC) techniques. They are also commonly used to characterize dynamic properties and structural changes upon local or environmental perturbations.

Structures deriving from folding or docking procedures need, in general, at least a structural regularization by EM before final validation steps, to avoid meaningless results from many methods. Scoring functions of the latter evaluate the probity of parameters, such as dihedral angle distributions, presence and distribution of steric bumps, voids in the molecular core, specific nonbonded interactions (H-bonds, hydrophobic clusters). Representing a mandatory step in most MC/MD protocols, EM programs are included in all the molecular simulation packages, and they share with MC/MD most input files and part of the setup parameters. Thus, unless they are be explicitly discussed, all system- and restraint-related features or issues illustrated for simulation methods also implicitly held for EM.

As we are mostly interested in techniques implementing experimentally-derived constraints or restraints, some of the most popular methods for constraints-based modelling will be briefly described. These methods have been developed and optimized mainly to identify and refine 3D structures consistent with spatial constraints from diffraction and resonance experiments (de Bakker et al., 2006). They have also been extensively applied to both TBM (Fiser & Sali, 2003) and free modelling prediction and simulation (Bradley et al., 2005; Schueler-Furman et al., 2005), and are often used to refine/validate models produced in TFM and TBM approaches described in sections 3.1 and 3.2. There are two main categories of constraint-based modelling algorithms: i) distance geometry embedding, which uses a metric matrix of distances from atomic coordinates to their collective centroid, to project distance space to 3D space (Havel et al. 1983; Aszodi et al. 1995, 1997); ii) minimization, which incorporates distance constraints in variable energy optimization procedures, such as molecular dynamics (MD) and Monte Carlo (MC). For both MD and MC, it is possible to work both in full cartesian coordinates, or in the restricted torsion angle (TA) space, with covalent structure parameters kept fixed at their reference values, thus originating the Torsional Angle MD (TAMD) and Torsional Angle MC (TAMC) approaches. They are currently implemented in several modelling and refinement packages, developed for structural refinement of X-ray or NMR structures (Rice & Brünger, 1994; Stein et al. 1997; Güntert et al., 1997), folding prediction (Gray et al., 2003), or more general packages (Mathiowetz et al., 1994; Vaidehi et al., 1997). Standard MC/MD methods are only useful for structural refinement, local exploration and to characterize limited global rearrangements. However, they are also widely used as sampling techniques in folding/docking approaches, although in those cases enhanced sampling extensions of both methods are employed. Simulated annealing (SA) (Kirkpatrick et al., 1983) and replica exchange (RE) approaches (Nymeyer et al., 2004) are the most common examples of these MC/MD enhancements, both potentially overcoming the large energy barriers required for sampling the wide conformational and configurational spaces to be explored in folding and docking applications, respectively.

A non-exhaustive list of the most diffused simulation packages including a more-than-basic treatment of distance-related restraints and also exhibiting good versatility (i.e. implementation of different algorithms, approaches, force fields and solvent representations), may include at least: AMBER (http://ambermd.org/), CHARMM (http://www.charmm.org/), DESMOND (http://deshawresearch.com/resources.html), GROMACS (http://www.gromacs.org/) and TINKER (http://dasher.wustl.edu/tinker). CYANA (http:www.cyana.org) and XPLOR/CNS (http://cns-online.org/v1.3/), although originally more specialized for structural determination and refinement from NMR and NMR/X-ray data, respectively, have been recently included in several TFM and TBM protocols, thanks to their efficient implementations of TAMD and distance or torsional angle restraints. The choice of a simulation program should ideally keep into account several criteria, ranging from computational efficiency, to support of sampling or refinement algorithms, to integration with other tools for TFM or TBM applications.

The main problems associated with simulation methods having relevant potential implications on MS3D are: i) insufficient sampling; ii) inaccuracy in the potential energy functionals driving the simulations; iii) influence of the approach used to implement experimentally-derived information on final structure sets.

Sampling problem can be approached both by increasing the sampling efficiency with MC/MD variations like SA and RE, and by decreasing the size of the space to be explored.

This latter result can be reached by reducing the overall number of degrees of freedom to be explicitly sampled and/or by reducing the number of possible values per variable to a small, finite number (discretization, like in grid-based methods), and/or by restraining acceptable variable ranges. Reduction of the total number of degrees of freedom can be accomplished by switching to coarse-grained representations of the system, where a number of explicit atoms, ranging from connected triples, to amino acid sidechains, to whole residues, up to full protein subdomains, are replaced by a single particle. This method is frequently used in initial stages of *ab initio* folding modelling, or in the simulation of very large systems, such as giant structural proteins of huge protein aggregates.

Another possible way to reduce the number of degrees of freedom is the aforementioned TA approach, requiring for a N atom system only N/3 torsional angles compared with 3N coordinates in atomic cartesian space (Schwieters & Clore, 2001). Moreover, as the high frequency motions of bending and stretching are removed, TAMD can use longer time steps in the numerical integration of equations of motion than that required for a classical molecular dynamics in cartesian space. Its main limitation may derive from neglecting covalent geometry variations (in particular, bending centred on protein Cα atoms) that are known to be associated with conformational variations (Berkholz et al., 2009), for instance from α-helix to β-strand, and that can be important in concerted transitions or in large structures with extensive and oriented SS regions. Discretization is mostly employed in the initial screening of computationally intensive problems, such as *ab initio* modelling. Restraining variable value ranges in MS3D is usually associated with either predictive methods (SS, H-bond pattern, residue exposure), or to homology analysis, or to experimentally-derived information. Origin, nature and form of these restraints have already been discussed in previous sections, while some more detail on the implementation of distance-related information into simulation programs will be given at the end of this section.

While the implementation of restraints can be very variable in methods where the scoring function does not intend to mimic or replicate a physical interaction between involved entities, in methods based on physically-sounding molecular potential functions (forcefields) have DRs implemented by a more limited number of approaches. At its simplest, a DR will be represented as a harmonic restraint, for which only the target distance and the force constant need to be specified in input. This functional form is present in practically all most common programs, but either requires a precise knowledge of the target distance, or it will result in a very loose restraint if the force constant is lowered too much to account for low-precision target values, the usual case in MS-based data. In a more complex and useful form, implemented with slight variations in several programs (AMBER, CHARMM, GROMACS, XPLOR/CNS, DESMOND, TINKER), the restraint is a well with a square bottom with parabolic sides out to a defined distance, and then linear beyond that on both (AMBER) or just the upper limit side (CHARMM, GROMACS, XPLOR/CNS, DESMOND). In some programs (CHARMM, AMBER, XPLOR/CNS), it is possible to select an alternative behaviour when a distance restraint gets very large (Nilges et al,1988b) by "flattening out" the potential, thus leading to no force for large violations; this allows for errors in constraint lists, but might tend to ignore constraints that should be included to pull a bad initial structure towards a more correct one.

Other forms for less-common applications can also be available in the programs or be implemented by an user. However, the most interesting additional features of versatile DR

implementations are the different averages that can be used to describe DRs: i) complex restraints can involve atom groups rather than single atoms at either or both restraint sides; ii) time-averaged DRs, where target values are satisfied on average within a given time lapse rather than instantaneously; iii) ambiguous DRs, averaged on different distance pairs. The latter two cases are very useful when the overall DRs are not fully consistent each other, because they are observed in the presence of conformational equilibria and, as such, they are associated with different microstates of the system. In addition, complex and versatile protocols can be simply developed in those programs where different parameters can be smoothly varied during the simulation (AMBER).

### 3.3.2 Programs for model validation

A validation of the final models, very often included in part in the available automated modelling protocols, represents a mandatory step, especially for more complex (low-homology, few experimental data) modelling tasks. A huge number of protein and nucleic acid structural analysis and validation tools exists, based on many different criteria, and subjected to continuous development and testing; thus, even a CASP section is dedicated to structural assessment tools (http://www.predictioncenter.org/), and the "Bioinformatics Links Directory" site alone currently reports 76 results matching "3-D structural analysis" (Brazas et al., 2010). Being outside the scope of the present report, information on 3D structural validation tools can be searched on specialized sites such as http://bioinformatics.ca/links_directory/ . However, similarly to what stated on prediction metaservers, a general principle for validation is to possibly use several tools, based on different criteria, looking for emergent properties and consensus among the results.

Specific parameters associated with MS-based data can be usually analysed with available tools. Distance restraints and their violations can be analysed both on single structures and on ensembles (sets of possible solutions of prediction methods, frames from molecular dynamics trajectories) with several graphic or textual programs, the most specialized obviously being those tools developed for the analysis of NMR-derived structures.

Surface information can be analysed by programs like:
DSSP (http://swift.cmbi.ru.nl/gv/dssp/),
NACCESS (http://www.bioinf.manchester.ac.uk/naccess/),
GETAREA (http://curie.utmb.edu/getarea.html/),
ASA-VIEW (http://gibk26.bse.kyutech.ac.jp/jouhou/shandar/netasa/asaview/)
that calculate different kinds of molecular surfaces, such as van der Waals, accessible, or solvent excluded surfaces for overall systems and contact surfaces for complexes are used. However, differently from distance restraints, available programs usually work on a single input structure at a time, thus making structure filtering and analysis on the large ensembles of models potentially produced by conformational prediction, molecular simulation or docking calculations, a painful or impossible task. In these cases, scripts or programs to automate the surface calculations and to average or filter the results must be developed.

## 4. Modelling with sparse experimental restraints

In the previous section many of the computational methods that can concur to produce structural models in MS3D applications have been outlined, together with different ways to integrate MS-based experimental information into them. Here we will refocus on the overall

computational approach in MS3D, to illustrate some of its peculiar features and issues, its present potentialities and the variety of possible combinations of data and protocols that can be devised to optimally handle different types of structural problems. Depending on nature and quantity of available experimental information and on previous knowledge of the investigated system, different combinations of the methods mentioned in previous sections can be optimally employed. We will start illustrating examples of methods for *de novo* protein folding, a frontier application of modelling with sparse restraints, because it is based on minimal additional information on the system under investigation.

The MONSSTER program (Skolnick et al. 1997) only makes use of SS profiles and a limited number of long-distance restraints. By employing system discretization and coarse-graining to reduce required sampling, a protein is represented by a lattice-based Cα-backbone trace with single interaction center rotamers for the side-chains. By using N/7 (N is the protein length) long-range restraints, this method is able to produce folds of moderate resolution, falling in the range from 4 to 5.5 Å of RMSD for Cα-traces with respect to the native conformation for all α and α/β proteins, whereas β-proteins require, for the same resolution, N/4 restraints. A more recent method for *de novo* protein modelling (Latek et al., 2007) adopts restrained folding simulations supported by SS predictions, reinforced with sparse experimental data. Authors focused on NMR chemical-shift-based restraints, but also sparse restraints from different sources can be employed. A significant improvement of model quality was already obtained by using a number of DRs equal to N/12.

As already stated by Latek and colleagues, the introduction of DRs in protein folding protocol represents a critical step that in principle could negatively affect the sampling of conformational space. In fact, restraint application at too early stages of calculations can trap the protein into local minima, where restraints are satisfied, but the native conformation is not reached. In addition to the number, even the specific distribution of long-range restraints along the sequence can affect the sampling efficiency. To test the influence of data sets in folding problem, we applied a well-tested protocol of SA, developed for AMBER program and mainly oriented to NMR structure determination, to the folding simulation of bovine pancreatic trypsin inhibitor (BPTI), by using different sets of ten long-distance restraints, randomly selected from available NMR data (Berndt et al., 1992), with optional inclusion of a SS profile. Fig. 3 shows representative structures for each restraint set.

The four native BPTI disulphide bridges were taken into account by additional distance and angle restraints. BPTI represents a typical benchmark for this kind of studies, due to its peculiar topology (an α/β fold with long connection loops, stabilized by disulphide bonds) still associated with a limited size (58 residues), and to the availability of both X-ray and NMR accurate structures. SA cycles of 50 structures each were obtained and compared for four combinations of three sets (S1-3) of ten long distance restraints, totally non-redundant among different sets and SS profiles: a) S1+SS profile; b) S1 alone; c) S2+SS profile; d) S3+SS profile. S1 set performed definitely better than the other two, its best model exhibiting a RMSD value of 2.4 Å on protein backbone of residues 3-53 from the representative NMR structure.

This set was also able to provide a reasonable low-resolution fold even in the absence of SS restraints (b). S3 resulted in almost correctly folded models, but with significantly worse RMSD values than S1 (c). In S3 pseudomirror images (d) of the BPTI fold occurred several times and only one model out of 50 was correctly folded (not shown).

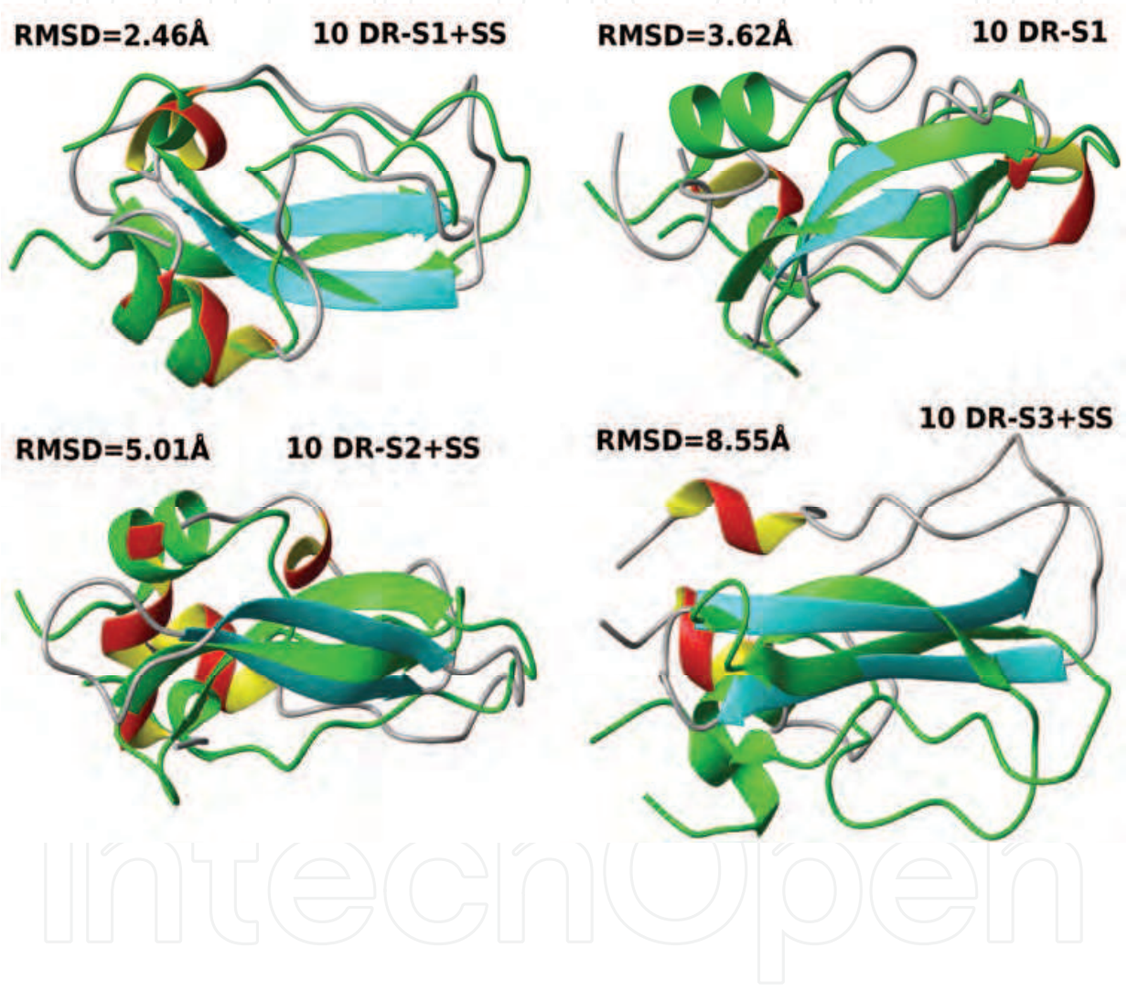



Fig. 3. *Ab initio* modelling from sparse restraints of BPTI. Representative models from different restraint sets (S1, S2, S3), with optional SS dihedral restraints are shown in ribbon representation, coloured in red/yellow (helices), cyan (strands) and grey (loops). Models are best-fitted on Cα atoms of residues 3-53 to the representative conformation of the NMR high-resolution structure (PDB code: 1pit) (green), except for the S3+SS set, where superposition with β-sheet only is shown, to better illustrate pseudo-mirroring effect, although RMSD values are calculated on the same 3-53 residue range as other models.

These results suggest a strong dependency of results upon both the exact nature of experimental data used in structure determination, and the protocol followed for model building. Thus, the number of restraints estimated in the aforementioned studies as necessary/sufficient for a reliable structural prediction should be prudently interpreted for practical purposes. If a proper protocol is adopted, increasing quantity, quality and distribution homogeneity of data should decrease this dependency, but the problem still remains severe when using very sparse restraints, such as those associated with many MS3D applications. A careful validation of the models and, possibly, execution of more modelling cycles with variations in different protocol parameters, can help to identify and solve these kinds of problems.

However, in spite of these potential issues, *ab initio* MS3D can provide a precious insight into systems that are impossible to study with other structural methods. In addition to increases in the number of experimental data, also homology-based information and other statistically-derived constraints can substantially increase the reliability of MS3D predictions. Thus, suitable combinations of experimental data, predictive methods and computational approaches have allowed the modelling of many different proteins and protein complexes spanning a wide range of sizes and complexity. The illustrative examples shown in Table 1 represents just a sample of systems affordable with current computational MS3D techniques and a guideline to select possible approaches for different problem classes. Heterogeneity of reported systems, data and methods, while suggesting the enormous potentialities of MS3D approaches, practically prevents any really meaningful critical comparison among methods, whose description in applicative papers is often incomplete. A standardization of MS3D computational methods is still far from being achieved, since it requires considerable computational effort to tackle with the considerable number of strategies and parameters that should be tested in a truly exhaustive analysis. Furthermore, the extreme sensitivity of modelling with sparse data to constraint distribution, as seen in the example shown in Fig. 3, either introduces some degree of arbitrariness in comparative analyses, or make them even more computationally-intensive, by requiring the use of more subsets for each system setup to be sampled.

Advancements in MS3D experimental approaches continuously change the scenarios for computational procedures, by substantially increasing the number of data, as well as the types of crosslinking or labelling agents and proteolytic enzymes. The large number of crosslinks obtained for apolipoproteins (Silva et al., 2005; Tubb et al., 2008) or CopA copper ATPase (Lübben et al., 2009) represent good examples of these trends (Table 1).

## 5. Conclusion

As already stated in the preceding section, the compared analysis of computational approaches involved in MS3D is still considerably limited, because of the complexity both of the systems to be investigated, and of the methods themselves, especially when they are used in combination with restraints as sparse as those usually available in MS3D studies. The continuous development in all involved experimental and computational techniques considerably accelerates the obsolescence of the results provided by any accurate methodological analysis, thus representing a further disincentive to these usually very time consuming studies. In this view, rather than strict prescriptions, detailed recipes or sharp critical compared analysis of available approaches, this study was meant to provide an

| System | Exp. data[a] | Use of exp. data[b] | Available info[c] | Additional data[d] | Modelling techniques[e]/ Refinement[e] | Software | Reference |
|---|---|---|---|---|---|---|---|
| Annexin A2/p11 complex | **3 XL, CL** | **XL**:**PSF**, **CL**:**PSF** | **XR**: ANXA2 (part); $(p11)_2$-(ANXA2 N-terminus)$_2$ complex | **SSp** for missing residues | multistep **PPD** / NA | Rosetta | Shultz et al., 2007 |
| Apolipoprotein (apo) A-I | **17 XL** | **XL**:**IIS** | **XR**: 4 templates | Helix amphi-patic profiles | **HM**( regions); **XL** / **EM** with **DR**s | MOE-AMBER | Silva et al., 2005 |
| Apolipoprotein (apo) A-IV | **21 XL** | **XL**:**IIS** | **XR**: res. 1-241: 1 template | **SSp:** residues 242-378 | **HM** of 1-241; merged with **SSp** by **XL** / **EM** with **DR**s | MOE, I-TASSER | Tubb et al., 2008 |
| Calmodulin (CaM)-Munc13 peptides complexes | **XL** (**6** x Munc13; **3**x ubMunc13-2); **PL** | **XL**,**PL**: **IIS**,**PSF** | **XR**: different CaM-peptides complexes | **SSp** for Munc13 peptides | **FP** of Munc13 peptides; multistep **PPD** with **IIS** & **PSF/NA** | Bhageerath program, Patch-Dock, Rosetta Dock | Dimova et al., 2009 |
| CopA copper ATPase | **18 XL** | **XL**:**IIS**, **PSF** | **XR**: different templates for the diverse domains | TM-helix predictions | **HM**; Multistep **PPD** with **PSF** and **IIS** / **EM** | SwissModel, 3Dgarden, Haddock, Xplor-NIH | Lübben et al., 2009 |
| Ffh-FtsY complex | **9 XL** | **XL**:**IIS** | **XR**: isolated proteins | NA | Multistep **PPD** with **IIS** / **SA-MD**, **EM** | CHARMM, Dock, Multidock | Chu et al., 2004 |
| G-actin–cofilin complex | **2 XL** | **XL**:**PSF** | **XR**: isolated proteins, templates for 5 N-ter residues of cofilin | NA | **HM** of missing regions; **PPD** with **PSF** / NA | InsightII, Autodock 3.0 | Grintsevich et al., 2008 |
| Latexin–carboxypeptidase A complex | **3 XL** | **XL**:**PSF** | **XR**: isolated proteins | NA | **PPD** with **PSF** / NA | In house rigid PD with hydrophobic energy score | Mouradov et al., 2006 |
| MutL - MutH complex | **4 XL** | **XL**:**PSF** | **XR**: isolated proteins | | **PPD** with **PSF** / NA | BIGGER | Giron-Monzon et al., 2004 |

www.intechopen.com

www.intechopen.com

| | | | | | |
|---|---|---|---|---|---|
| Urease complex | **2 XL** | **XL**:**PSF** | **XR**: template | NA | **HM** with **PSF / ** NA |
| GTPases Rab4 and Rab5 | **LP** | **LP**:**PSF** | **XR**: templates | NA | **HM** with **PSF / EM** |
| ERK2 - PTP-SL complex | **XL** | **XL**:**IIS** | **XR**: templates | NA | **HM**, **aiM**, **PPD** with **IIS / SA-MD**, **EM** |
| Aminoacylase 1 | **1 XL**, **CL**, **LP** | **XL**,**CL**: **PSF** | **XR**: templates | **SSp** | **HM** with **PSF / EM** |
| Sso7d-melittin / ATP complex | **LP**, **PL** | **LP**:**PSF**, **PL**: **IIS** | **XR**: isolated proteins | NA | **PPD**, **PLD** with **IIS** & **PSF / MD**, **EM** |
| Gadd45β-MKK7 complex | **LP**, **AA** | **LP**,**AA**: **IIS** | **XR**: isolated proteins | **SSp** | **HM** with **IIS**, **PPD** with **PSF/MD,EM** |
| Calmodulin-Melittin complex | **LP,CL** | **LP,CL: PSF** | **XR**: CaM, CaM-peptides complexes | **NA** | **MD/EM** |

Table 1. Some examples of MS3D studies from literature. The following abbreviations have been use
abbreviations): [a]**XL**: crosslinking, **CL**: chemical labelling, **PL**: photoaffinity labeling, **LP**: limited prot
analyses, [b]see [a], **PSF**: post-sampling filtering with experimental data, **IIS**: esperimental data integrat
crystallography; [d]**SSp**: secondary structure prediction, [e]see [a,b], **HM**: Homology modeling, **aiM***: ab-in*
prediction, **MD**: molecular dynamics, **SA**: Simulate Annealing, **EM**: energy minimization, **PPD**: pro
protein-small ligand docking; **DR**: distance restraints, TM: trans-membrane; NA: not available.

overall and as wide as possible picture of the state-of-art approaches in MS3D computational techniques and their potential application fields. However, in spite of these limitations, some general conclusions can still be drawn.

For predictive methods that stay behind the most ambitious MS3D applications (*ab initio* folding, folding prediction, threading), at least when used in the absence of experimental data, metaservers exhibit on average best performances than the single employed servers, as also shown by the results of the last CASP rounds on automatic servers (http://predictioncenter.org/). This suggests two distinct considerations: 1) the accuracy of sampling and scoring exhibited by each single method, as well the rationale behind them, are still so limited to prevent reliable predictions on best performing methods in any given case; 2) nevertheless, most methods tend to locate correct solutions, or, in general, groups of solutions including the correct one or a close analogue. Therefore, a consensus among the predictions from different servers generally improves the final solutions, by smoothing down both extreme results and random fluctuations associated with each single approach. Well consolidated metaservers, such as Robetta or I-TASSER, can be regarded as reasonable starting guesses for general folding problems, also considering that they both include distance-related restraints in their available options. However, special classes of systems (e.g. transmembrane proteins or several enzyme families) can instead benefit from employing specifically-devised approaches.

In comparing server-based applications to standalone programs (often available in alternative for a given approach), potential users should also consider that the former require less computational skill and resources, but are intrinsically less flexible than the latter, and that legal and secrecy issues may arise, because several servers consider submitted prediction requests and the corresponding results as public data, as usually clearly stated in submission pages. In addition to possible information "leakage" in projects, the public status of the models would prevent their use in patents.

When considering more specifically MS3D procedures, it has been shown that even a small number of MS-based restraints can significantly help in restricting the overall space to be explored and in identifying the correct fold/complexation mode, especially if they are introduced in early modelling stages of a computational procedure optimized to deal with both the investigated system and the available data. Thus, experimental restraints can allow the use of a single model generation procedure, rather than a multiple/metaserver approach, at least in non-critical cases. In fact, they should filter out all wrong solutions deriving from the biases of the modelling method, leaving only those close to the "real" one, if it is included in the sampled set. In particular, since the lowest energy structure should ideally also be associated with a minimum violation of experimentally-derived restraints, the coincidence of minimum energy structures with least violated restraints should be suggestive of correct modelling convergence and evaluation of experimental data. However, particular care must be adopted not only in the choice of the overall computational procedure, but especially of the protocol used to introduce experimental information, because a too abrupt build up of the restraints can easily bring to local minima far from the correct solution. Comparison of proper scoring functions other than energy between experimentally-restrained and unrestrained solutions may provide significant help in identifying potential issues in data or protocols. Estimates of the sensitivity of solutions to changes in protocols may also enforce the reliability of best converged cases. In particular, when other restraints are also present, the relative strength and/or introduction order of the

different sets could play an important role in the final result; thus, their weight should be carefully evaluated by performing more modelling runs with different setups.
When evaluating the overall modelling procedures, their corresponding caveats and performance issues, the importance of many details in setup and validation of MS3D computational procedures fully emerges, thus suggesting that they still requires a considerable human skill, although many full automated programs and servers allow in principle the use of MS3D protocols even to inexperienced users. This is also demonstrated for the pure *ab initio* modelling stage by the still superior performances obtained by human-guided predictions in CASP rounds, when compared to fully automated servers.
Future improvements in MS3D are expected as a natural consequence of continuous development in biochemical/MS techniques for experimental data, and in hardware/ software for molecular simulations and predictive methods. However, some specific, less expensive and, possibly, quicker evolution in MS3D could be propelled by targeted development of computational approaches more directly related to the real nature of the experimental information on which MS3D is based, notably algorithms implementing surface-dependent contributions and more faithful representations of crosslinkers than straight distance restraints.

## 6. References

Aebersold, R. & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, Vol.422, pp. 198–207.

Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, Vol.25, pp.3389–3402.

Aszodi, A.; Gradwell, M.J. & Taylor, W.R.(1995). Global fold determination from a small number of distance restraints. *Journal of Molecular Biology* Vol.251, pp.308–326.

Aszodi, A.; Munro, R.E. & Taylor, W.R.(1997). Protein modeling by multiple sequence threading and distance geometry. *Proteins*, Vol. 29, pp.38–42.

Back, J.W.; de Jong, L.; Muijsers, A.O. & de Koster, C.G. (2003). Chemical crosslinking and mass spectrometry for protein structural modeling. *Journal of Molecular Biology,* Vol.331,pp.303–313.

Back, J.W.; Sanz, M.A.; De Jong, L.; De Koning, L.J.; Nijtmans, L.G.; De Koster, C.G.; Grivell, L.A.; Van Der Spek, H. & Muijsers, A.O.(2002). A structure for the yeast prohibitin complex: Structure prediction and evidence from chemical crosslinking and mass spectrometry. *Protein Science*, Vol. 11, pp.2471–2478.

Balasu, M.C.; Spiridon, L.N.; Miron, S. ; Craescu, C.T.; Scheidig, A.J., Petrescu, A.J. & Szedlacsek, S.E. (2009). Interface Analysis of the Complex between ERK2 and PTP-SL. *Plos one*, Vol. 4, pp. e5432.

Bastard, K.; Prévost, C. & Zacharias, M. (2006). Accounting for loop flexibility during protein-protein docking. *Proteins,* Vol.62, pp. 956-969.

Ben-Zeev, E. & Eisenstein, M. (2003). Weighted geometric docking: incorporating external information in the rotation-translation scan. *Proteins,* Vol.52, pp. 24-27.

Berndt, K.D.; Güntert, P.; Orbons, L.P. & Wüthrich, K. (1992). Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic

trypsin inhibitor and comparison with three crystal structures. *Journal of Molecular Biology,* Vol.227, pp.757-775.

Blake, J.D. & Cohen, F.E. (2001). Pairwise sequence alignment below the twilight zone. *Journal of Molecular Biology*, Vol. 307, pp. 721-735.

Bowers, P.M.; Strauss, C.E.M. & Baker, D. (2000). De novo protein structure determination using sparse NMR data. *Journal of Biomolecular NMR*, Vol.18, pp.311–318.

Brazas, M.D.; Yamada, J.T. & Ouellette, B.F.F. (2010). Providing web servers and training in Bioinformatics: 2010 update on the Bioinformatics Links Directory . *Nucleic Acids Research*, Vol. 38, pp.W3–W6.

Brooks, B.R.; Bruccoleri, R.E.; Olafson, B.D.; States, D.J.; Swaminathan, S. & Karplus M.(2003). CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry*, Vol.4, pp.187-217.

Camacho, C. J. & Vajda, S. (2001). Protein docking along smooth association pathways. *PNAS USA,* Vol.98, pp.10636–10641.

Carlsohn, E.; Ångström, J. ; Emmett, M.R.; Marshall, A.G. & Nilsson, C.L. (2004). Chemical cross-linking of the urease complex from Helicobacter pylori and analysis by Fourier transform ion cyclotron resonance mass spectrometry and molecular modeling *International Journal of Mass Spectrometry*, Vol.234, pp. 137–144.

Chu, F.; Shan, S.; Moustakas, D.T.; Alber, F.; Egea, P.F.; Stroud, R.M.; Walter, P. & Burlingame A.L. (2004). Unraveling the interface of signal recognition particle and its receptor by using chemical cross-linking and tandem mass spectrometry. *PNAS*, Vol.101, pp. 16454-16459.

D'Ambrosio, C.; Talamo, F.; Vitale, R.M.; Amodeo, P.; Tell, G.; Ferrara, L. & Scaloni, A. (2003). Probing the Dimeric Structure of Porcine Aminoacylase 1 by Mass Spectrometric and Modeling Procedures. *Biochemistry*, Vol. 42, pp. 4430-4443.

de Bakker, P.I.; Furnham, N.; Blundell, T.L. & DePristo, M.A. (2006). Conformer generation under restraints. *Current Opinion in Structural Biology*, Vol. 16, pp.160–165.

Dimova, K; Kalkhof, S.; Pottratz, I.; Ihling, C.; Rodriguez-Castaneda, F.; Liepold, T.; Griesinger, C.; Brose, N.; Sinz, A. & Jahn, O. (2009). Structural Insights into the Calmodulin-Munc13 Interaction Obtained by Cross-Linking and Mass Spectrometry. *Biochemistry*, Vol.48, pp. 5908-5921.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics,* Vol.14, pp.755–763.

Fabris, D. & Yu, E.T. (2010). The collaboratory for MS3D:a new cyberinfrastructure for the structural elucidation of biological macromolecules and their assemblies using mass spectrometry-based approaches. *Journal Proteome Research*, Vol.7, pp. 4848-4857.

Fiser, A. & Sali, A. (2003). Modeller: generation and refinement of homology base protein structure models. *Methods in Enzymology*, Vol. 374, pp.461–491.

Förster, F.; Webb, B.; Krukenberg, K.A.; Tsuruta, H.; Agard, D.A. & Sali A.(2008). Integration of Small-Angle X-Ray Scattering Data into Structural Modeling of Proteins and Their Assemblies. *Journal of Molecular Biology,* Vol.382, pp.1089–1106.

Friedhoff, P. (2005). Mapping protein–protein interactions by bioinformatics and crosslinking.. *Analitycal & Bioanalitycal Chemistry,* Vol.381,pp.78–80.

Giron-Monzon, L.; Manelyte, L.; Ahrends, R.; Kirsch, D.; Spengler, B. & Friedhoff, P. (2004). Mapping Protein-Protein Interactions between MutL and MutH by Cross-linking. *The Journal of Biochemical Chemistry*, Vol.279, pp. 49338–49345.

Gray, J.J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C.A. & Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology,* Vol.331, pp.281-299.

Green, N.S.; Reisler, E. & Houk, K.N. (2001). Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers. *Protein Science,* Vol.10, pp.1293-1304.

Grintsevich, E.E.; Benchaar, S.A.; Warshaviak, D.; Boontheung, P.; Halgand, F.; Whitelegge, J.P.; Faull, K.F.; Ogorzalek Loo, R.R; Sept, D.; Loo, J.A. & Reisler, E. (2008). Mapping the Cofilin Binding Site on Yeast G-Actin by Chemical Cross-Linking. *Journal of Molecular Biology*, Vol.377, pp. 395-409.

Güntert, P.; Mumenthaler, C. & Wüthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program Dyana. *Journal of Molecular Biology*, Vol. 273, pp. 283–298.

Havel, T.F.; Kuntz, I.D. & Crippen, G.M.(1983). The combinatorial distance geometry method for the calculation of molecular conformation. I. A new approach to an old problem. *Journal of Theoretical Biology*, Vol. 310, pp.638–642.

Jaroszewski, L.; Rychlewski, L.; Li, Z.; Li, W. & Godzik, A. (2005). FFAS03: a server for profile– profile sequence alignments. *Nucleic Acids Research*, Vol.33, pp.W284–288.

Karplus, K.; Barrett, C. & Hughey R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, Vol.14, pp.846–856.

Kessl, J.J.; Eidahl, J.O.; Shkriabai, N.; Zhao, Z.; McKee, C.J.; Hess, S.; Burke, T.R. Jr & Kvaratskhelia, M. (2009). An allosteric mechanism for inhibiting HIV-1 integrase with a small molecule.*Molecular Pharmacology*, Vol. 76, pp.824–832.

Kirkpatrick, S.; Gelatt, C.D. Jr. & Vecchi, M.P. (1983). Optimization by Simulated Annealing. *Science*, Vol 220,pp. 671-680.

Latek, D.; Ekonomiuk, D. & Kolinski , A.(2007). Protein structure prediction: combining de novo modeling with sparse experimental data. *Journal of Computational Chemistry*, Vol. 28, pp.1668–1676.

Leitner, A.; Walzthoeni, T.; Kahraman, A.; Herzog, F.; Rinner, O.; Beck, M. & Aebersolda, R. (2010). Probing Native Protein Structures by Chemical Cross-linking, Mass Spectrometry, and Bioinformatics. *Molecular & Cellular Proteomics*, Vol.24, pp. 1634-1649.

Lin, M.; Lu, H.M.; Rong Chen,R. & Liang, J.(2008). Generating properly weighted ensemble of conformations of proteins from sparse or indirect distance constraints. *The Journal of Chemical Physics,* Vol.129, pp.094101–094114.

Marti-Renom, M.A.; Madhusudhan, M.S. & Sali, A. (2004). Alignment of protein sequences by their profiles. *Protein Science*, Vol.13, pp.1071–1087.

Lübben, M.; Portmann, R.; Kock, G.; Stoll, R.; Young, M.M. & Solioz, M. (2009). Structural model of the CopA copper ATPase of Enterococcus hirae based on chemical cross-linking. *Biometals*, Vol.22, pp. 363-375.

Mathiowetz, A.M.; Jain, A.; Karasawa, N. & Goddard, W.A. III. (1994). Protein simulation using techniques suitable for very large systems: The cell multipole method for nonbond interactions and the Newton–Euler inverse mass operator method for internal coordinate dynamics. *Proteins*, Vol. 20, pp. 227–247.

Melo, F. & Sali, A. (2007). Fold assessment for comparative protein structure modeling. *Protein Science*, Vol. 16, pp. 2412–2426.

Millevoi, S.; Thion, L.; Joseph, G.; Vossen, C.; Ghisolfi-Nieto, L. & Erard, M. (2001). Atypical binding of the neuronal POU protein N-Oct3 to noncanonical DNA targets. Implications for heterodimerization with HNF-3b. *European Journal Biochemistry*, Vol.268, pp. 781-791.

Moreira, I.S.; Fernandes, P.A. & Ramos, M.J. (2010). Protein-protein docking dealing with the unknown. *Journal of Computational Chemistry*, Vol. 31, pp.317–342.

Mouradov, D.; Craven, A.; Forwood, J.K.; Flanagan, J.U.; García-Castellanos, R.; Gomis-Rüth, F.X.; Hume, D.A.; Martin, J.L.; Kobe, B. & Huber, T. (2006). Modelling the structure of latexin–carboxypeptidase. A complex based on chemical cross-linking and molecular docking. *Protein Engineering, Design & Selection*, Vol.19, pp. 9-16.

Nikolova, L.; Soman, K. ; Nichols, J.C.; Daniel, D.S., Dickey, B.F. & Hoffenberg, S. (1998). Conformationally variable Rab protein surface regions mapped by limited proteolysis and homology modelling. *Biochemical Journal*, Vol.336, pp. 461–469.

Nilges, M.; Clore, G.M. & Gronenborn, A.M.(1988a). Determination of three dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Letters,* Vol.229, pp.317–324.

Nilges, M.; Gronenborn, A.M.; Brünger, A.T. & Clore, G.M. (1988b). Determination of three- dimensional structures of proteins by simulated annealing with interproton distance restraints: application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Engineering*, Vol.2, pp.27-38.

Nymeyer, H.; Gnanakaran, S. and García, A.E. (2004). Atomic simulations of protein folding using the replica exchange algorithm. *Methods in Enzymology,* Vol.383, pp.111-149.

Papa, S.; Monti, S.M.; Vitale, R.M.; Bubici, C.; Jayawardena, S.; Alvarez, K.; De Smaele, E.; Dathan, N.; Pedone, C.; Ruvo M. & Franzoso, G. (2007). Insights into the structural basis of the GADD45beta-mediated inactivation of the JNK kinase, MKK7/JNKK2.. *Journal of Biological Chemistry*, Vol. 282, pp. 19029-19041.

Potluri, S.; Khan, A.A.; Kuzminykh, A.; Bujnicki, J.M., Friedman, A.M. & Bailey-Kellogg, C. (2004). Geometric Analysis of Cross-Linkability for Protein Fold Discrimination. *Pacific Symposium on Biocomputing*, Vol.9, pp.447-458.

Renzone, G.; Salzano, A.M.; Arena, S.; D'Ambrosio, C. & Scaloni, A.(2007a). Mass Spectrometry-Based Approaches for Structural Studies on Protein Complexes at Low-Resolution. *Current Proteomics*, Vol. 4, pp. 1-16.

Renzone, G.; Vitale, R.M.; Scaloni, A.; Rossi, M., Amodeo, P. & Guagliardi A. (2007b). Structural Characterization of the Functional Regions in the Archaeal Protein Sso7d. *Proteins: Structure, Function, and Bioinformatics*, Vol. 67, pp. 189-197.

Rice, L.M & Brünger, A.T. (1994). Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins*, Vol. 19, pp. 277–290.

Russell, R.B.; Alber, F.; Aloy, P.; Davis, F.P.; Korkin, D.;Pichaud, M; Topf, M. & Sali, A. (2004). A structural perspective on protein-protein interactions. *Current Opinion in Structural Biology*, Vol.14, pp. 313-324.

Scaloni, A; Miraglia, N.; Orrù, S.; Amodeo, P.; Motta, A.; Marino, G. & Pucci, P.(1998). Topology of the calmodulin-melittin complex. *Journal of Molecular Biology*, Vol. 277, pp.945–958.

Schrag, J.D.; Jiralerspong, S.; Banville, M; Jaramillo, M.L. & O'Connor-McCourt, M.D. (2007). The crystal structure and dimerization interface of GADD45gamma. *PNAS*, Vol. 105, pp. 6566-6571.

Schueler-Furman, O.; Wang, C.; Bradley, P.; Misura, K. & Baker, D. (2005). Progress in modeling of protein structures and interactions. *Science*, Vol. 310, pp.638–642.

Schulz,D.M.; Kalkhof, S.; Schmidt, A.; Ihling, C.; Stingl, C.; Mechtler, K.; Zschörnig, O & Sinz, A. (2007). Annexin A2/P11 interaction: New insights into annexin A2 tetramer structure by chemical crosslinking, high-resolution mass spectrometry, and computational modeling. *Proteins: Structure Function & Bioinformatics*, Vol.69, pp. 254-269.

Schwieters, C.D. & Clore, G.M. (2001). Internal Coordinates for Molecular Dynamics and Minimization in Structure Determination and Refinement. *Journal of Magnetic Resonance*, Vol. 152, pp.288–302.

Silva, R.A.G.D.; Hilliard, G.M.; Fang, J.; Macha, S. & Davidson, W.S. (2005). A Three-Dimensional Molecular Model of Lipid-Free Apolipoprotein A-I Determined by Cross-Linking/Mass Spectrometry and Sequence Threading. *Biochemistry*, Vol.44, pp. 2759-2769.

Singh, P.; Panchaud, A. & Goodlett, D.R. (2010) Chemical Cross-Linking and Mass Spectrometry As a Low-Resolution Protein Structure Determination Technique. *Analytical Chemistry*, Vol. 82, pp. 2636–2642

Sinz, A. (2006). Chemical cross-linking and mass spectrometry to map three dimensional protein structures and protein-protein interactions. *Mass Spectrometry Reviews*, Vol.25, pp. 663-682.

Skolnick,J.; Kolinski, A. & Ortiz, A.R.(1997). MONSSTER: A Method for Folding Globular Proteins with a Small Number of Distance Restraints. *Journal of Molecular Biology*, Vol. 265 pp.217-241.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, Vol.21, pp.951–960.

Stein, E.G.; Rice, L.M & Brünger, A.T. (1997). Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *Journal of Magnetic Resonance*, Vol. 124, pp. 154–164.

Tubb, M.R.; Silva, R.A.G.D.; Fang, J.; Tso, P. & Davidson, W.S. (2008). A Three-dimensional Homology Model of Lipid-free Apolipoprotein A-IV Using Cross-linking and Mass Spectrometry. *The Journal of Biochemical Chemistry*, Vol.283, pp. 17314--17323.

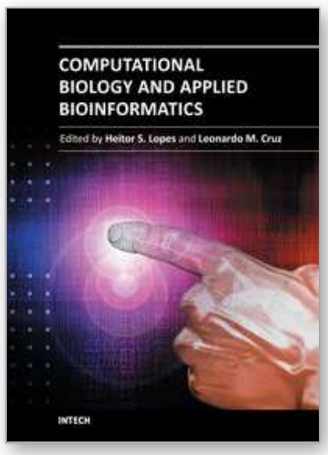
Vaidehi, N., Jain, A. & Goddard, W.A. III (1996). Constant temperature constrained molecular dynamics: The Newton–Euler inverse mass operator method. *Journal of Physical Chemistry*, Vol. 100, pp.10 508–10517.

Van Dijk, A.D.J.; Boelens, R. & Bonvin, A.M.J.J. (2005). Data-driven docking for the study of biomolecular complexes. *FEBS Journal,* Vol.272, pp.293–312.

Young, M.M.; Tang, N.; Hempel, J.C.; Oshiro, C.M.; Taylor, E.W.; Kuntz, I.D.; Gibson, B.W. & Dollinger G. (2000). High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *PNAS*, Vol.97, pp. 5802-2806.

Zheng, X.; Wintrode, P.L. & Chance M.R. (2007). Complementary Structural Mass Spectrometry Techniques Reveal Local Dynamics in Functionally Important Regions of a Metastable Serpin. *Structure*, Vol.16, pp. 38-51.

IntechOpen




**Computational Biology and Applied Bioinformatics**

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4

Hard cover, 442 pages

**Publisher** InTech

**Published online** 02, September, 2011

**Published in print edition** September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.

**How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Rosa M. Vitale, Giovanni Renzone, Andrea Scaloni and Pietro Amodeo (2011). Computational Methods in Mass Spectrometry-Based Protein 3D Studies, Computational Biology and Applied Bioinformatics, Prof. Heitor Lopes (Ed.), ISBN: 978-953-307-629-4, InTech, Available from: http://www.intechopen.com/books/computational-biology-and-applied-bioinformatics/computational-methods-in-mass-spectrometry-based-protein-3d-studies

INTECH
open science | open minds

**InTech Europe**

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166

**InTech China**

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

www.intechopen.com

IntechOpen

IntechOpen

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen