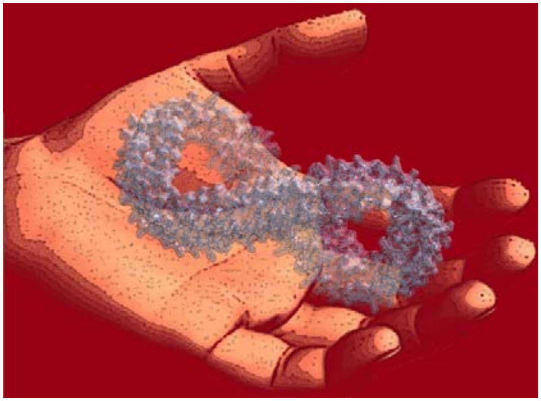
Università degli Studi di Roma "La Sapienza"

"Resource for benchmarking the applicability of protein structure models".



Tutore
Prof.ssa Anna Tramontano

Docente guida
Prof.ssa Anna Tramontano

Coordinatore
Prof. Marco Tripodi

Dottorando
Daniel Carbajo Pedrosa

Dottorato di Ricerca in Scienze Pasteuriane
XXIV CICLO

A mis padres, mi hermano, Carmen, Tita e Ión.

Abuelos no os olvido.

SUMMARY

1 - INTRODUCTION

1.1 - From structure to function

The function of a protein is closely related to the structure it attains (or lack of stable structure in the case of intrinsically disordered proteins). In enzymes, it is the specific spatial orientation of the residues configuring the catalytic sites what brings about their functionality; besides, and also true for transport proteins, enzyme structures need to accommodate to geometrical and electrostatic elements of the substrates and ligands to bind them specifically. Clearly the shape of proteins is also fundamental for constructing different cellular components. Thus, the sequence of a protein is of limited biological relevance without some knowledge of both its structure and its function; protein structures provide a wealth of information that cannot be deduced from their primary sequence alone; therefore, we can get a complete understanding of protein roles by analyzing them in structural terms.

Structure-based methodologies are consequently regarded as more robust than sequence-based ones. In particular, trying to predict the function of a protein relying only on sequence signatures, even if such information is sometimes critical, is considered insufficient, for the simple fact that the catalytic function of an enzyme, for example, is dictated by its 3D structure, which is often more stringently conserved than its sequence along evolution, diverging slower. This is because amino acidic residues not strictly necessary for the chemistry of the protein activity eventually disappear as sequences diverge, while the structure in the 3D space is retained. However, there are cases when biologically relevant residues are so stringently conserved at the sequence level that structural information does not provide an additional value; an example of these cases will be detailed in RESULTS.

1.2 - Need for protein structure models

There is no doubt that the prediction of the function of a protein can be helped by the knowledge of its structure (or lack of stable structure). The limiting step is actually having a protein's structure at hand; there is a relatively small number of experimentally determined protein structures, stored in the PDB (the Protein DataBank) [1], compared to the number of known protein sequences [2] (the difference is a factor of a thousand) (Figure 1).

A large amount of protein sequence data is being produced by large-scale DNA sequencing projects (such as the Human Genome Project) [3] [4] [5]; however, the output of experimentally-determined protein structures is lagging far behind, despite the ongoing efforts in structural genomics, mainly due to limitations of current structure determination techniques. These techniques, typically X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy, require high

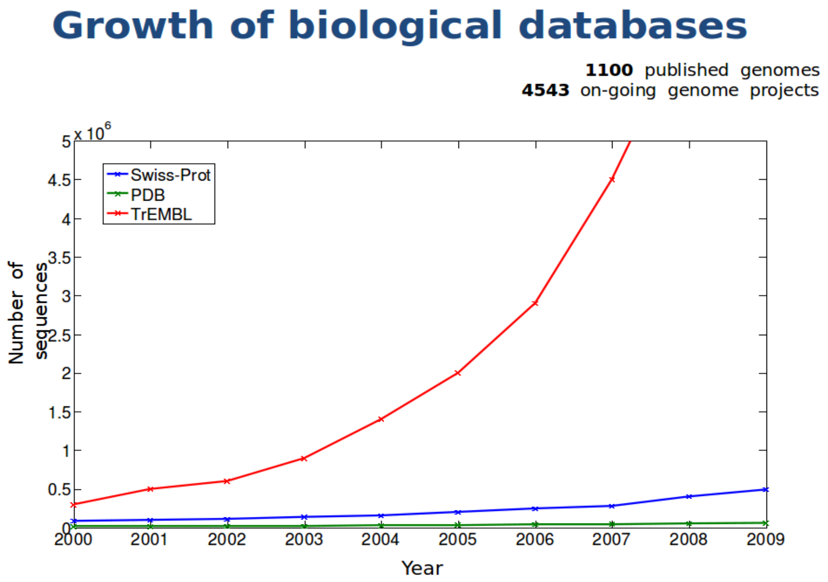


Figure 1: *Protein sequences deposited in TrEMBL, containing unreviewed, automatically annotated entries, in red. Protein sequences stored in Swiss-Prot, containing reviewed, manually annotated entries, in blue. Experimentally determined protein structures in PDB, in green. There is a large gap between known protein structures and sequences, and a larger gap between automatically and manually annotated sequences.*

level of protein expression (X-ray), a determined protein size (NMR), highly specialized equipment, staff and long execution times. There is no hope that the structure of all the around 3 million known proteins can be experimentally determined in the foreseeable future. As a result, the use of protein structure models is often necessary; scientists need to increasingly rely on them to understand the function of a protein with no known structure.

Inferring the native structure of a given protein (or the lack of a stable structure) should only demand the knowledge of its amino acidic sequence and of the environmental conditions. Nevertheless, protein structure prediction is still an extremely difficult task to tackle, being regarded as one of the biggest problem in current biological research.

1.3 - Protein structure prediction methods

Computational approaches exist for predicting unknown protein 3D structures from known primary amino acid sequences, helping to bridge the ever-increasing gap between known protein sequences and 3D structures. These can be classified as template-free and template-based approaches. Templates are proteins with experimentally determined structures that can be used as frameworks for building structural models of other proteins, called targets.

1.3.1 - Template-free prediction methods

One of the template-free approaches, *ab initio* (or *de novo*) *in silico* protein structure prediction, which will not be described in depth here, relies directly on general physical laws that govern protein folding energetics, using as the only target-specific information the protein sequence under study. This approach has been considered the "holy grail" of computational biology for the last 40 years.

Ab initio methods are limited to relatively small and topologically simple structures, mainly owing to the fact that current computers are not powerful enough to face the conformational sampling problem; the free energy landscape, or possible conformations a protein can adopt, that must be searched is astronomically large (even small proteins have on the order of 1000 degrees of freedom). These methods thus tend to demand vast computational resources; in order to predict the structure of a large and/or complex protein, more accurate algorithms and more powerful computers are needed.

Another non-template-based strategy, in the sense that it does not attempt to detect templates, for tertiary structure prediction is known as fragment assembly. As its name indicates, it is based on the assembly of structural fragments (for example, compact 3D structures of adjacent secondary structure elements, smaller than a domain or a subunit) subsequently optimized using a simulated annealing algorithm, and aims at narrowing the search of the conformational space of a protein by preselecting fragments from a library of resolved protein structures. This approach first splits the target protein sequence into small fragments, each of which is used to search corresponding structural fragments in a library (rather than using the entire target protein to search a protein structure library), identified based on energetic and geometric criteria; the full atom model is obtained assembling the structures of the segments. Fragment-based methods can be combined with template-based ones to model variable regions, such as loops, once the protein core has been built.

Though considered not mature enough, template-free methodologies need to be used for a good amount of protein sequences for which there is no structurally similar protein, or it cannot be identified by current tools, allowing the identification of new folds; in these cases, the comparative or template-based approach, referred to as the most efficient and described next, cannot be applied.

1.3.2 - Template-based prediction methods

These methods (also called comparative modeling) assume that the target protein folds in a similar way to that of a template protein (or a combination of template proteins) with a experimentally determined structure available, and are considered to perform more effectively.

Homology modeling requires that the target and the template are homologous, i.e. evolutionary related. 25 years ago Chothia and Lesk determined that naturally occurring homologous proteins usually have similar stable tertiary structures [6] [7], being more similar when proteins are more closely related (Figure

2). However, it is important to note here that there are known exceptions and nearly identical amino acid sequences do not always fold similarly [8] [9]; conformations may differ significantly due to environmental conditions as well, and intrinsically disordered proteins represent an extreme example.

Fold recognition methods, on the other side, aim at detecting analogous proteins with no evolutionary relationships, but with a similar fold. They rely almost exclusively on structural information, using statistical potentials to evaluate the fit between the target sequence and a template structure. These methods can be applied because the structural space of possible protein conformations (a limited set of tertiary structural motifs

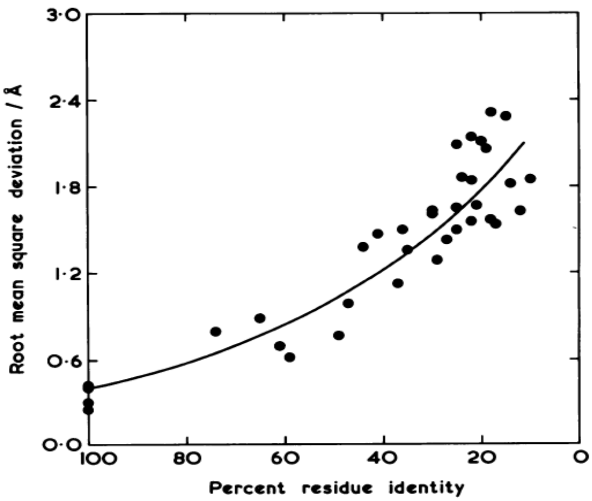


Figure 2: *Relation between sequence identity and structural similarity (in terms of Root Mean Square Deviation -RMSD- explained in METHODS) in different pairs of homologous proteins. The more similar proteins are at the sequence level, the more similar they will be at the structural level. Extracted from [7].*

to which most proteins belong) is much more restricted compared to the amount of all possible sequences (it has been suggested that there are no more than 2,000 unique folds in nature, while there are millions of proteins) [10]. In fact, up to 90% of the structures deposited in the PDB have similar folds to ones already there (according to the CATH protein structure classification [11] release notes). Thus, applying fold recognition methods, potential templates for a target protein can be found when they are structurally similar to the target, even if their sequences have diverged beyond recognition.

Fold recognition and homology modeling differ only at the template selection step; homology modeling can be applied for targets with homologous proteins with known structure, fold recognition is useful for other targets with only fold-level similarity.

1.3.2.1 - Steps in template-based prediction methods

A multi-step iterative process demanding the use of different tools needs to be performed to obtain a protein model by comparative methods. The steps are: Detection of templates with known 3D structures, selection of suitable templates, target-template(s) alignment, model building and model refinement. An additional step, namely model quality assessment (MQA), closely linked to the whole comparative modeling process, is used to evaluate whether a new iteration is required and/or a certain step polished.

Template detection

Protein templates with known 3D structures are essentially detected by searching a database of experimentally determined structures, such as the PDB or a subset of it, using the target protein as query.

The most traditional and simplest algorithms, named sequence-sequence comparison, align target sequences to those in the protein structure database in a pairwise fashion. Programs belonging to this class of simple sequence searches, such as BLAST (Basic Local Alignment Search Tool) [12], return the most similar protein sequences in the database (hits or potential templates) given a protein sequence query, but they may miss distant relatives due to low sensitivity [13] [14] [15].

More sensitivity can be achieved by building sequence profiles, also called position-specific scoring matrix (PSSMs) or position-specific weight matrix (PSWM), representing a complete protein family. They are matrices of scores with one row for each of the 20 amino acids and one column for each position in the query sequence; each score is derived from each amino acid frequency in each position and it reflects its evolutionary conservation and importance in defining a family member. Sequence profiles can be used as query against a sequence database, as is the case of PSI-BLAST (Position-Specific Iterative BLAST) [12], useful for finding distant homology.

Whereas sequence-sequence and profile-sequence comparison methods perform satisfactory when the sequence identity between the target and the potential templates is above 40%, when identity drops below 30% (near the so-called twilight zone and beyond), they are not as sensitive as profile-profile and HMM-HMM comparison methods. Profile-profile approaches are able to detect distant homology more effectively by constructing a sequence profile for the target query protein and comparing it to pre-calculated profiles for the potential templates with known 3D structure in the database being searched [16]. The use of HMMs (Hidden Markov Models) further increases sensitivity and specificity; HMMs also include position-specific amino acid frequencies for insertions and deletions (indels) in a probabilistic model, making comparison methods that use them able to build accurate alignments [17] [18].

Structural information can be integrated into profiles and HMMs by adding new vectors or dimensions to the sequence-only derived matrix or new states to the HMM. Tools that use HMMs combined with a structural trait (generally secondary structure), are called hybrid (they cannot be classified as homology modeling or fold recognition) and have been proved to outperform the above mentioned methods.

The best example of those arguably most sensitive hybrid tools to date [19]

[20] [21] [22], which make use of HMMs combined with structural information, is the one applied in the present project, HHsearch, also available as an on-line version called HHpred [20]. HHsearch, as described in METHODS, is a popular free software package useful for identifying more remote homologous proteins or protein families, when no relatives can be found using more conventional comparison methods.

Regardless of the method used, once potential templates with known structures are identified, the most suitable one(s) have to be selected.

Template selection and target-template(s) alignment

Selecting proper templates for building a model has a direct impact on the quality of the obtained model, since similar folds can differ from each other in local details. Potential templates are commonly ranked by the search methods described above, according to various scores. The template(s) selected should be the most structurally similar to the target protein of unknown structure. It is advantageous to use different search methods as most suitable templates might be detected by a consensus strategy. If no consensus exists, however, a careful selection must be carried out by the user based on all available information.

One can assume that templates with higher sequence similarity (closer in evolution) compared with the target sequence might have a more similar structure.

Other knowledge can be applied when selecting proper templates; for instance, a template should be selected if its structure was determined under the same environmental conditions as the ones required for the structural model, e.g. a ligand-bound conformation, or if it has a better resolution than others, or lower temperature factor (which indicate the fluctuation of an atom about its average position, as explained below), or does not have missing coordinates.

Once a proper template, or more than one, has been selected, a sequence alignment with the target protein has to be built; including several family members in a multiple sequence alignment has been shown to be more effective [23] [24] [25] [26]. A target sequence can be accurately modeled on distantly related templates with known structure, provided that the relationships between them and the target can be detected in the alignment. Although producing precise models remains a challenge when no close homologs are available, it has been suggested that sequence alignment is the bottleneck in the process [27]. Unsurprisingly, homology modeling is most reliable when the target and template(s) have rather similar sequences.

Comparative search methods often provide single template-target alignments as outputs; while those produced by HMM-based tools are particularly accurate, they should be manually refined if a precise model is to be built. The use of different search methods and the comparison of the alignments they provide might help to identify regions aligned with high confidence; one should also bear in mind that hydrophobic residues should be kept in the protein interior and indels should be placed in solvent-exposed regions, without defined secondary structure (usually loops), not interfering with biologically relevant sites (active or binding), expected to be more stringently conserved throughout evolution, accumulating fewer changes [6] [28] [29].

Model building and refinement

Having a target-template(s) alignment and known experimentally determined structures for the templates, a 3D structural model for the target protein, represented as a set of 3D Cartesian coordinates for each atom in the protein, can be constructed in an automatic fashion using any of the current tools devised for that purpose. The alignment specifies which residues of the target should be modeled following spatial specifications of the corresponding residues in the template(s) structure(s).

Spatial restraints satisfaction is a model building technique, inspired by NMR, that automatically calculate the model for the target protein, containing all non-hydrogen atoms, that best fulfills various constraints derived from the template(s) and mapped onto each target residues following the alignment provided. Stereochemical constraints like bond lengths and angles, planarity of peptide groups and side-chain rings, chiralities, van der Waals contact distances, bond angles, dihedral angles etc. are encoded in a probability density function that the model has to minimize. The main representative of this strategy is Modeller, the tool selected for our research, which has been shown to outperform many other programs [30] [31] [32] [33].

Even models produced by state-of-the-art programs are suboptimal [34] and should therefore be refined as a whole or focusing on particular regions; the refinement, a step not taken in our approach, and therefore not discussed here, is usually carried out based on physical or statistical potentials.

Loop and side-chain modeling, other steps not discussed here, are also crucial for obtaining a full-atom model. Loops are regions not stringently conserved throughout evolution, where indels are more frequent, important to model as they might contain biologically relevant sites (they are flexible and thus suitable for accommodating a ligand); loops are usually modeled based either on the conformation of loops with known structures extracted from a library, on physical and knowledge-based potentials to select the most correct conformation between a series of alternatives (something computationally demanding for long loops), or on hybrid methods that combine the previous database search and *ab initio* ones. Side-chains, on the other side, are usually constructed by many modeling methods, but can be rebuilt and added to the backbone of the final model; methods like Modeller explicitly build these side-chains by applying torsion angle information derived from the templates or from rotamer libraries.

1.3.3 - Model quality assessment (MQA)

This step, closely linked to the modeling process (either template-free or template-based) defines whether the process should undergo a further iteration, polishing one or more of the described steps. MQA programs try to establish the overall correctness of a model obtained following the above described procedures, or local accuracies of small portions of it; this, in turn, determines the model usefulness for specific applications and the information that can be extracted from it [30] [35]. Given a poor score for a final comparative model, one should go back to previous steps and either select different template(s), make modifications in the alignment, use a different modeling program or refine the obtained model. Hence, it is advantageous to generate alternative versions of the alignment with different templates (extracted from different search methods) and different modifications in order to assess the different models obtained and select the best one. This is straightforward using the so-called metapredictors or metaservers, like Pcons [36] [37] and 3D-Jury (a simplified version of Pcons) [38], which either automatically collect the models computed by several predictor servers and rank them, or combine the best predicted local portions of the different models returning a single one; these metapredictors represent the best current solution for the protein structure prediction issue.

It is also particularly important to measure the accuracy of certain regions of the model independently since the entire model is unlikely to be modeled with the same accuracy; if the structural quality of a certain region cannot be improved, one should be aware when trying to use the model for a specific application.

While sequence identity is quite a good indicator of the overall correctness of a model when sequence identity is above 40% [39], it becomes unreliable when sequence identity drops below 30%, when alignment errors become frequent, and MQA programs are essential (these cases, in fact, are the most common [40]). These programs fall into two different categories depending on how the accuracy score is derived: The first comprises those programs that derive the score of a single model from evolutionary [41] [42] [43] or physicochemical and statistical criteria, by for example measuring how common are the model geometrical features with respect to those of known well-defined high-resolution structures [44] [45] [46] [47] [48] [49] [50] [51]; multivariate model assessment methods, like GenThreader [52] or the Modeller-8 program (where the authors found that the score mainly depends on residue accessibilities and distances, model compactness and percentage of sequence identity between target and template(s) used in the alignment [53]) also belong to this category. The second category is represented by those other programs, designated as clustering or consensus programs, such as Pcons and 3D-Jury (which also incorporate MQA) that derive the score out of the information contained in an ensemble of alternative models one usually ends up with while attempting to predict the structure of a target protein, via an all-against-all comparison, based on the assumption that the more frequently a conformation is predicted, the more likely it is correct [38] [54] [55]. These approaches, although shown to be more consistent, are only useful when a large set of models with significant structural diversity for the same target are available, which is frequently the case using current procedures, especially metapredictors.

Recently developed MQA servers, like QMEAN [56], offer the possibility to use scoring functions of both of the described categories, and both locally and globally.

Depending on the quality of the final model obtained, it could be used for different purposes, discussed below.

1.3.4 - Assessment of protein structure prediction methods

No matter the modeling method, but especially true for comparative ones, the quality of the predicted structures has improved, as measured by blind tests in the CASP (Critical Assessment of Methods of Protein Structure Prediction) series of meetings, where the classical observation of Chothia and Lesk that evolutionary related proteins almost always have similar 3D structure has been repeatedly confirmed [57]. CASP is an on-going community-wide experiment where the state-of-the-art protein structure prediction methods are assessed on a two-year basis, since 1994 [58].

Blind targets, without an available structure (either structures soon-to-be solved by experimentalists or already solved but kept on hold by the PDB), are proposed, so none of the participants starts from an advantageous position. Once the experimental native structure of the target protein is available, it is used as the "gold standard" to evaluate the submitted models. Finding the optimal superposition between model and native structure is not a trivial task; the correctness of a superposition is directly correlated with the fraction of the structures that is superimposed [59] (structurally similar regions might be obviated if whole structures are considered for the superposition). Detecting well-predicted regions has a biological implication, as they might correspond to relevant features of the protein studied.

GDT-TS (Global Distance Test - Total Score) is nowadays the standard evaluation measure of model correctness in the CASP experiment [60]. It is regarded as more suitable than RMSD (Root Mean Square Deviation), as explained below.

Determining which protein structure method works better, the final aim of the CASP experiment, is difficult, since not all methods are suitable for all targets and targets range in relative difficulty. In order to come up with a method ranking (as well as to define the method improvement over the two-year gap between CASP rounds), the evaluation panel needs to consider how many and which targets it predicts correctly. The difficulty of a target can be defined *a posteriori*, averaging the correctness of the different predicted models using the different methods, but it can also be defined *a priori*, considering how problematic it can be to identify evolutionary relationships with templates used, if any.

1.4 - Importance of protein structure prediction, remaining challenges and perspectives

As stated above, compared to the about 3 million public available protein sequences resulting from methodological advances in DNA sequencing, there are relatively few experimentally determined 3D structures and the use of models is considered mandatory in many scenarios. Even if experimental structures have been determined for only 1% of all the identified proteins, reliable models can be computed for up to 20%. Structure prediction methods are not likely to replace experimental determination of structures, but rather to complement them; biologists can use computed models to guide experimental design.

Comparative or template-based modeling is regarded as the most effective methodology, yielding more accurate models than *ab initio* or template-free modeling, which has not experienced as much progress [61] [62] [63]. Furthermore, as the number of deposited experimental structures increases (following improvements in current experimental methods X-ray crystallization and NMR), structure prediction approaches will eventually be restricted to comparative modeling, as templates representing most of the protein families will become available. Models obtained will therefore improve; nowadays, however, a predicted template-based model is seldom found to be significantly closer to the target native structure than the template used.

It seems then, that we are in the correct path, but there is still a lot ahead for comparative approaches to be able to reproduce native-like folds of target proteins. After years of development, detecting remotely homologous templates (with still structural similarity) and building accurate alignments still remain the two issues having the major impact on the quality of resulting models. When sequence identity between template(s) and target is higher than 40%, though, alignment discrepancies are rare, so the main focus is to model and refine accurately variable regions (indels) and side chains; in these cases, models generated are considered comparable in quality to native structures solved by low resolution X-ray crystallography [63] [64] [35].

As already commented, the structural space is restricted, so as more experimentally determined protein structures are deposited in public databases and more reliable models can be obtained for the rest, following the improvement observed especially in template-based modeling strategies, we will get closer to have already seen all possible distinct folds in nature. Exploring the complete space of protein structure is, finally, within our reach.

1.5 - Predicted protein models applications

The amount of information that can be derived from a model ultimately depends on its quality, thus, as models are nothing but predictions, one has to know the corresponding estimates of model quality before using them for a certain purpose. Potential biological applications with regard to model quality are commented in [65] summarized in Figure 3.

High-quality models, typically those with an RMSD value of 1-2Å with respect to the native counterpart and usually constructed by template-based protocols when sequence identity between target and template(s) is higher than 40%, are functionally interpretable in general, as alignments are rarely incorrect, backbone

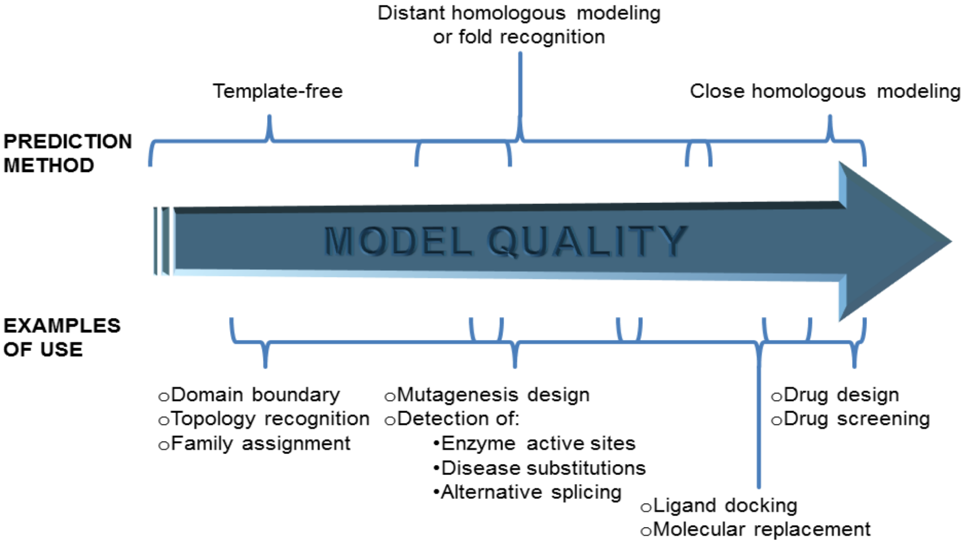


Figure 3: *Approximate correspondence of the accuracy and the biological usefulness of protein structure predictions. Modified from [65].*

conformation is expected to be close to the native, and many side-chains are usually correctly oriented [30] [66]. If the templates used are functionally characterized, they can be used to infer the function of the target proteins; if they are not, the models obtained still serve as good starting points for detailed bioinformatics function predictions. The level of detail obtained with these models is sometimes sufficient even for drug design. Some researchers have successfully used computed protein structure models to guide the design of new drugs [67] [68]. The range of applications is wide, from enzyme reaction mechanisms inferences [69] and disease-causing mutations interpretation [70], to computational ligand-docking studies [71] (which predict the preferred orientation of one molecule to a second when bound to each other to form a stable complex [72]) and even experimental structure determination aid [73] [74].

Mid-range quality models, roughly considered those with an RMSD of 2-5Å and usually obtained by homology modeling with distant relatives or by fold recognition, might be helpful as well in delineating the spatial locations of biologically relevant sites, such as active sites or disease-associated mutations. As will be explained below, groups like Arakaki's one have attempted to detect the quality cutoff of structural models that still allows the assignment of an enzymatic function by matching structural patterns of active sites; this group concluded that models in the RMSD range of 3-4Å from the experimental native structure are sufficient for the transfer of the first three EC (Enzyme Commission) numbers (a numerical classification scheme for enzymes based on the chemical reactions they catalyze [75]) [69].

Low accuracy models (say, with an RMSD higher than 3Å with respect to the native counterpart), usually obtained applying template-free modeling strategies, on the other hand, should not be used to answer "high-resolution" questions, but can find other applications; having even an inaccurate structural model at hand can significantly facilitate the handling of a protein in diverse experiments. Models of approximately correct fold can be used to recognize the overall topology, to identify domain boundaries [76] [77] or to infer functionalities at least up to a family level [78] [79]. Functional analysis of these models is not easy, because functional regions might be incorrectly built up by, for instance, misplaced secondary structure elements with incorrect side-chain orientations. However, modifications occur at a slower rate in functional sites, so even distantly related proteins can have functional sites with similar geometry (structure tends to be more stringently conserved than sequence) [6] [28] [29]. Thus, if the target and a template bear similar activity, the structure of the functional site might be modeled correctly even if the overall sequence identity is low.

1.6 - Aim of our study

Central to the problem of assessing the quality of protein structure prediction methods is the question of how good should a model be in order to be used in place of its corresponding experimentally determined structure for a given structure-based method. In other words, even if we know what is the expected error of a model on the basis of its sequence similarity with a homologous protein of known structure, we still have no clear idea about how this numerical estimate translates into biological usefulness. For example, if it can be estimated that a given model is, on average, 1Å away from the real structure, there is at present no clear idea, despite some attempts, about whether such error is compatible with, say, drug design or docking of the macromolecular structural model with its partner(s). We believe that this is of fundamental importance if the models have to be used in a real setting by the biological community.

Our approach here has consisted in developing a system that builds sets of models of decreasing quality, which we call decoys, given the sequences experimentally determined proteins. A decoy is a computer-generated protein structure that possesses some characteristics of native proteins, but is not biologically real. Our system is implemented in such a way that any structure-based existing method can be tested on the real structure and on the decoy models. The next step is to automatically assess at which level of quality the results of the tested method differ from those obtained with the native structure.

Thus, the tool we developed, named ModelDB, publicly available either as an on-line tool or a local application for larger calculations, is intended to provide a series of decoy homology models out of any protein structure. A resulting decoy set will be composed by the query native structure plus a variable number of decoy models of different quality. The quality of a decoy models is determined by the amount and severity of structural inaccuracies it includes, which are left unrefined, and is assessed comparing it directly to the native structure; these inaccuracies can be wrong overall folds due to the selection of a non suitable template, wrong residue spatial positions due to misalignments, displacement of local secondary structure elements, a trend in protein evolution not uncommon amongst distant homologs used as templates, misfolded loops and insertions with no counterpart in the template, and wrong side-chain conformations.

As a significant subset of the PDB is modeled, all the decoy models generated represent a significant portion of the protein structural space, and cover all possible accuracies and possible applications; each decoy set taken individually will cover a quality range depending on the templates found for the target query experimentally determined native PDB structure**.**

Decoy models are primarily meant to be used to test structure-based methods, such as a functional site predictor that employs 3D features shown by such biologically relevant sites. Our aim is to provide a public tool that is able to build a decoy set out of any input protein; making use of decoy sets generated with our tool, one can benchmark the applicability of a given structure-based method to computed models, making it feasible to derive a model quality tolerance for such method.

Hence, we aim at addressing the following question: To what extent do models allow for a reliable structural analysis? How good should the models be?

For a visual insight of how models of different qualities look like and differ from the native counterpart in a spatial context, the on-line version of the tool we developed incorporates a visualization window where the user can manipulate the different structures he/she wants to load. Biologically relevant sites are highlighted for the native structure and models, so one can study the extent they distort as model overall quality decreases.

Focusing on these relevant sites (in particular active sites), we also aim at understanding whether they can be identified in models of decreasing quality, if they are preserved in low quality models, and to what extent can these models be used to study functional sites in the protein structural space. Another future goal would be to determine whether the conservation/distortion of functional sites specific and distinctive structural characteristics across models of decreasing quality can be used as a means of evaluating a model.

1.7 - Bases of our study

Being the models created based on templates, the present method has its roots based on two well-known and efficient free tools in the field of comparative modeling, HHsearch coupled with Modeller. Models are constructed by Modeller based on the atomic coordinates of single templates detected by HHsearch and selected when fulfilling thresholds that make the range of model quality span the maximum possible still allowing each model to cover the maximum extent of the native structure.

Providing a quality measure for each decoy model generated is clearly important if one is to derive a quality cutoff for a structure-based method, or just to know the extent of information that can be inferred from a given model, as we have seen. Each model is directly compared to its corresponding native structure and precise quality scores are computed.

Furthermore, for a visual insight on how models of different qualities look like and differ from the native counterpart in a spatial context, they are "colored" following different color schemes defined by the following spatial descriptors: Solvent accessibilities, secondary structures, cavity occurrences, average depths, protrusion indexes or burial indexes. This, in turn, allows an easy visualization and understanding of these parameters' variations in the protein structural context.

Besides, functional annotation is provided when available, in terms of catalytic sites, ligand-binding sites and other sites of relevance like glycosylation sites.

1.8 - Examples of previous uses of decoy structures

Decoys are often used in protein folding studies to test the validity of a protein model; the model will be considered correct if it is able to identify the native state configuration of the protein among the decoys. In the same sense, the primary use of decoys is to test and improve MQA scoring or energy functions, determining whether the native structure or native-like structures can be found in a large set of decoys (examples can be found in the literature: [80] [81] [82] [83] [84] [85] [86] [87] [88] [89] [90] [91] [92] [93]).

Many applications for a computed model in a given resolution range were verified by the use of sets of decoy models of varying quality. These applications, explained next, range from locating biologically relevant sites, to simulating protein docking, to help experimental determination of protein structures.

A recent use of decoy structures was to measure to which extent computed protein structure models can be used in ligand screening simulations; Brylinski and Skolnick developed and improved Q-Dock [71], a low-resolution flexible ligand docking approach able to recover 62-87% of binding residues using distorted receptor structures with an RMSD up to 3Å, 25-35% more (and 15-20% more specific contacts) than all-atom methods, proved to be not tolerant enough to structural deformations of the ligand binding region [94] [95] [96]. Q-Dock can deal with these deformations as it uses a low-resolution coarse-grained docking approach (it considers fewer and larger components than fine-grained all-atom approaches). Other low-resolution docking techniques are listed in references [97] [98] [99] [100].

Another application restricted to high-quality models and evidenced by the use of decoys (in particular those submitted to the CASP experiment) is that of experimental structure determination aid. This is accomplished by means of molecular replacement, a technique where an initial model is used to solve the phasing problem, a major one in X-ray crystallography. The performance of this technique depends on the global, rather than local, quality of the initial model, which has to have a GDT-TS higher than 84% with respect to the native structure (the ranking in terms of RMSD is more blurred); different initial models were found to be significantly more successful than the raw templates used to build them in molecular replacement experiments, highlighting the importance of structural refinement [73] [74]. More recently, it was demonstrated that not only template-based high-resolution models, but also high-resolution models refined from NMR structures and even from template-free models can be successfully used for molecular replacement [101].

Taylor developed a method for comparing two protein structures based on decoys. Such method is able to measure the significance of a comparison in way independent from their nature and their level of representation in the PDB; decoys for the two proteins being compared are computed in order to be used as a background score distribution. Not relying on homologous structures deposited in the PDB, this method is suitable for comparing membrane proteins, RNA structures, or distorted structures like *ab initio* models [102].

As the elucidation of function is one of the main intentions of structure prediction, other researchers have taken advantage of decoy structures motivated by the need to determine the model resolution tolerance of their own structure-based function prediction algorithms. Most of those are successful when applied to high-resolution structures, but only a few cases have been tested for lower resolution predicted structures [103] [104].

Fetrow and Skolnick presented a method for the identification of protein function that relies on the sequence-to-structure-to-function paradigm; they developed descriptors for active sites, termed "fuzzy functional forms", based on their conformation and geometry. They showed that these "fuzzy functional forms" can identify protein active sites not only from experimentally evidenced structures, but also from computed ones provided by either *ab initio* or by fold recognition algorithms, proving that low-to-moderate resolution structures are sufficient to identify enzymatic active sites [103]. However, their conclusions are not applicable when functional sites, either enzymatic or binding, such as calcium-binding ones, lack strongly conserved residues or residue geometry [105] [106].

Wei *et al.* questioned the utility of computed protein models for the identification of calcium-binding sites. They used decoy structures stored in the Decoys 'R' Us database [107] for the vitamin D dependent calcium-binding protein to test their method. They did not find a correlation between the overall RMSD (Root Mean Square Deviation) of a structure and the calcium-binding site microenvironment preservation. The microenvironments of calcium-binding sites are reliably modeled only in high-quality structures; however, the overall structure quality is very weakly related to its functional sites microenvironments quality, since functional sites tend to be local entities rather than global ones [104]. Nevertheless, they suggest that the selection of the correct protein fold from a large amount of well predicted decoy structures might be easier if they were filtered using functional site-recognition methods.

Arakaki and coworkers tested their own method more recently with decoy structures [108]. They attempted to predict the enzymatic active sites of given proteins in a structure-based fashion, generating libraries of functional 3D descriptors, termed "automated functional templates" [69] (which take into account information extracted from public databases, such as Swiss-Prot and PDB), and adding this information to annotation transfer from the given proteins' homologs [109] [110]. Homology-based inference routines alone tend to fail as the sequence identity trespasses the twilight zone, and they can lead to errors due to the functional promiscuity shown by many protein families [111] [112]. There is still controversy around the sequence identity threshold for these routines to succeed, in terms of EC numbers transfer, [113] [114] [115] [116] [112] [117], but most researchers agree that 40% sequence identity is enough to transfer the first three EC components with an estimated accuracy of 90%. By adding structural information, Arakaki's group was able to increase this accuracy. Matching the structural patterns of active sites with decoys of different resolution, they concluded that a given enzyme biochemical function (transfer of at least the first three EC numbers) could be correctly recognized in 35% of the cases using models in the range of 3-4Å RMSD away from the native structure, something that can be readily achieved using even current *ab initio* methodologies [118] [119].

Chelliah and Taylor have also recently tested their previously developed methodology called CRESCENDO, useful for predicting residues likely to compose an active or binding site [120], applying it to a large collection of decoy models obtained from what is known as the "Periodic Table" classification of protein structures [121]. They verified that their method obviously performed much better on native-like folds, suggesting, yet again, that filtering of structures having well formed functional sites can indeed help selecting the best structures among a set of predicted ones [122].

1.9 - Original contribution of our study and potential applications

Due to the ever-increasing gap between known protein sequences and structures and the ever-growing number of protein structure prediction methods available, which are becoming more and more accurate over time, the use of protein structure models is mandatory, specially if one is to predict the function of proteins without a known structure. However, and in spite of progress in the field of protein structure prediction, computed models often contain structural inaccuracies in both backbone and side-chain spatial coordinates (it is quite unlikely that models fall near the free energy neighborhood of the native structure); instead of being discarded, these models can provide important insights into the function of the native counterpart; this, in turn, demands the existence of robust methods that can effectively make use of computed models in the midrange and low range of accuracy, routinely produced by proteome-scale protein structure modeling projects. Any structure-based predictor, such as an active site predictor, that does not require high-resolution structures for the successful identification of biologically relevant sites, as well as any other structure-based algorithm, will prove to have a big advantage and an inestimable practical value. These predictors, in particular, could allow us to extend functional analyses into and beyond the twilight zone of sequence identity, something relevant given the fast expansion of genomics databases.

ModelDB, the tool introduced here, strives to serve as a resource to test any structure-based method (such as an active site or ligand-binding site predictor) on protein structure models of different quality. This has the final goal of benchmarking the applicability of a given novel algorithm to protein structure models.

Very few other public resources exist for readily retrieving decoy sets of protein structures, and we indeed have no record of any other automated pipeline for producing such decoys in an easy and user-friendly fashion. The main resource to date on the web to retrieve decoys would be Decoys 'R' Us [107]; another main decoy set source is CASP, where researchers can easily gather all the submitted models for any of the targets. Nonetheless our tool, apart from allowing to build new decoy sets for a given protein a scientist is interested in, covers many more different proteins representing a larger portion of the protein structural space. Furthermore, the on-line version has the advantage to let the user visually inspect and compare all the models of ranging quality for a given protein in the same spatial frame; the functional documentation, the model quality estimates and the structures color schemes following different spatial descriptors allow users to effectively examine how biologically relevant sites distort as model quality drops.

2 - METHODS

2.1 - Starting dataset

Several decoy sets and modified decoy sets (as explained below) have been already created out of a subset of PDB structures. The subset comprises those protein and protein-nucleic acid complexes solved by X-ray crystallography as of 3rd January 2011, culled at 50% sequence identity, excluding those structures with only Cα atoms, those with a resolution worse than 2Å (the lower the resolution the better is the quality of the structure), those with a sequence length outside the range 20-10000 residues, and those with an R-factor higher than 0.3 (a quality measure defining the fit between the refined crystallographic model and the observed experimental X-ray diffraction data; a perfect fit would have a value of 0), for a total of 8,609 PDB chains. This was accomplished making use of PISCES, useful for filtering PDB lists in order to obtain the longest possible subset meeting given sequence identity and structural quality criteria [123].

2.2 - Modeling strategy

The ModelDB modeling pipeline, written in Perl, relies on two state-of-the-art algorithms in the field of template-based modeling: HHsearch and Modeller, and is used to generate a decoy model per each of the PDB chains in the input list. The steps, detailed next, are summarized in Figure 4.

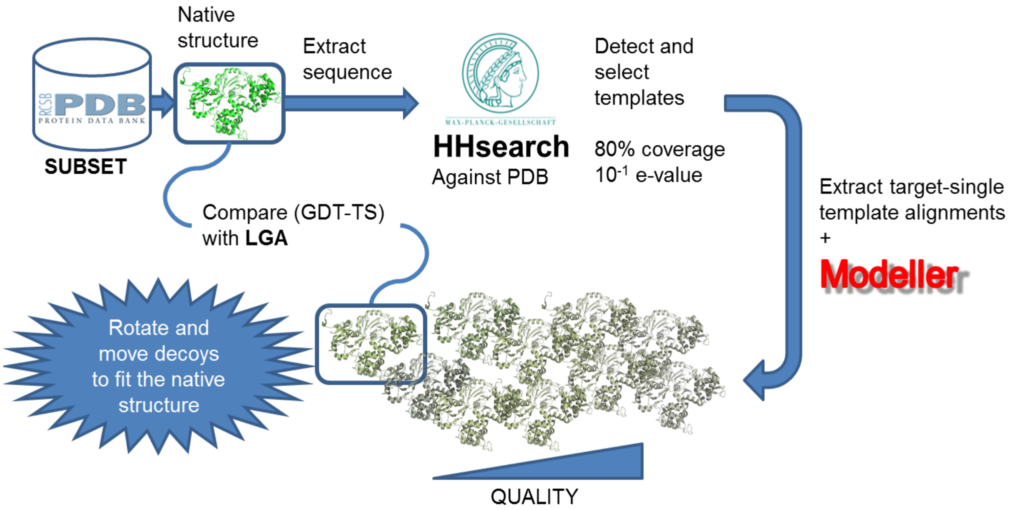

Native structure
RCSB PDB PROTEIN DATA BANK
SUBSET
Extract sequence
HHsearch
Against PDB
Detect and select templates
80% coverage
10^{-1} e-value
Extract target-single template alignments
+
Modeller
Compare (GDT-TS) with LGA
Rotate and move decoys to fit the native structure
QUALITY

Figure 4: *Modelling pipeline: Given a target protein structure (pre-calculated decoy sets exits for a subset of the PDB), its sequence is used as query for HHsearch to search for templates in PDB; target-single template alignments are extracted and used as guidelines for building models of different qualities using Modeller; the qualities are measured by comparing each model obtained with the target native structure using LGA; each model is moved and rotated to fit the native structure.*

For the sake of time-saving, sequences, instead of HMMs, were used as queries for HHsearch to search for templates with known experimental structure in the 70% non redundant PDB HMM database. Target-single template alignments were extracted when potential templates fulfilled the thresholds of 80% minimum coverage (aligned residues divided by the query residue length and multiplied by 100) and 10^{-1} maximum e-value (the average expected number of non-homologous proteins with a score higher than the one obtained for the database match).

For each target, Modeller was applied to produce an all non-hydrogen atom single-template model for each of the templates selected, that best satisfies restraints derived from such template, using the alignment extracted from HHsearch as guideline.

2.3 - Native structure - models superimposition and models classification

A precise quality estimation for each model can be provided since native structures are available. Every decoy set is composed by the native query structure and a series of models of decreasing quality in terms of GDT-TS and RMSD as calculated by LGA (Local-Global Alignment) [124] between each model and the native structure. As implemented in the ModelDB procedure, LGA serves for comparing two proteins or portions of them at the structural level. The LGA GDT algorithm is designed to search for the largest (not necessarily continuous) set of "equivalent" residues deviating by no more than a specified distance cutoff.

Thus, every decoy is supplied along with its quality as compared with the native query structure. GDT-TS is regarded as more suitable than RMSD, as it portrays the number of Cα atom pairs of the model and the target native structure that are close enough to derive meaningful insights from the model (i.e. the percentage of those Cα atoms in the model falling within a defined distance cutoff of 1, 2, 4 and 8Å from their native position), following the formula:

$$GDT-TS=100*\frac{\sum_{d=1}\frac{GDT_{di}}{NT}}{4} \qquad d_i \in \{1.0, 2.0, 4.0, 8.0\}$$

It is the average of the four GDT scores obtained using the four distance (*d*) cutoffs (*i* equal to 1, 2, 4 or 8Å), divided by the number of residues of the target native structure (*NT*).

RMSD, a quadratic measure, is the square root of the squared differences between atom pairs in model and native structure, following the formula:

$$RMSD=\sqrt{\frac{1}{N}\sum_{i=1}^{N}[(x_i-x'_i)^2+(y_i-y'_i)^2+(z_i-z'_i)^2]}$$

Where (x_i, y_i, z_i) and (x'_i, y'_i, z'_i) are the atomic coordinates in the 3D

Cartesian space of one structure (native) and the other (model), respectively and N the number of atom pairs compared. This way, regions not correctly superimposed weight more. Especially in the group of low resolution models, RMSD becomes a non informative measure, as local misorientation of tails and loops may result in a big overall RMSD that can mask a correctly modeled core region, for example. In these cases, the use of GDT-TS becomes especially important.

Other structural quality estimates apart from GDT-TS and RMSD and extracted from LGA are LGA_S and LGA_Q, defined in reference [124]. Estimates derived from HHsearch pairwise alignments between the target native protein sequence and each of the templates used to build the single-template models are also provided; these include percentage sequence identity, e-value, coverage, probability and HHsearch score.

LGA's supplied rotation and translation matrices are used to rotate and move each decoy model coordinate in the Euclidean space so as to best fit the native structure, allowing for a spatial superimposition of all the structures in the decoy set, using any standard molecular visualization system, like PyMol [125] or Jmol [126].

All these decoy sets generated constitute the pre-calculated results of the public web server described below. Each model in a decoy set can be easily ranked according to the different estimates listed.

2.4 - Functional annotation

Each PDB structure in the input list (as well as its decoys) is functionally annotated when possible, using three main sources: The CREDO database [127], the Catalytic Site Atlas (CSA) [128] [129], and Swiss-Prot.

CREDO is a comprehensive relational database documenting protein-ligand interactions derived from structures stored in the PDB. CSA stores information about enzyme active sites and catalytic residues, whenever an enzyme has a 3D structure available. Only those residues thought to be directly involved in some aspect of the reaction catalyzed by an enzyme are considered, constituting two types of entries: Hand-annotated ones, extracted from the literature, and homologous ones, derived via PSI-BLAST.

Other functional or biologically relevant annotation is retrieved from Swiss-Prot, well known for being a high-quality, manually curated, non redundant protein sequence database, conceived to provide all known relevant information about a protein. Swiss-Prot features considered functional or biologically relevant are the ones in the following list: active site, binding site for any chemical group, glycosylation site, calcium-binding region, disulfide bond, DNA-binding region, domain, intramembrane region, covalent binding of a lipid moiety, binding site for a metal ion, posttranslational modification of a residue, short sequence motif, mutagenesis site, nucleotide phosphate-binding region, region of interest, signal sequence, interesting non-defined site, topological domain, transmembrane region and zinc finger region.

These CREDO, CSA and Swiss-Prot features are mapped onto corresponding structures by an in-house bl2seq-based Perl-BioPerl program called MAP. Bl2seq uses the same BLAST algorithm to compare one sequence to another.

2.5 - Structure coloring

For a visual insight on how models of different qualities look like and differ from the native counterpart in a spatial context, each structure in a decoy set is "colored". This coloring is achieved by replacing the temperature factor records in the file containing the atomic coordinates of each structure in the decoy set with a residue or atom-based spatial descriptor out of the following: Solvent accessibilities, secondary structures, cavity occurrences, average depths, protrusion indexes or burial indexes. This allows an easy visualization of these parameters' value distributions in space within a structure and amongst structures of increasingly lower quality, providing insight on the spatial behavior of each residue, specially those biologically relevant. This is achieved thanks to three different programs: DSSP (Define Secondary Structure of Proteins) [130], Speedfill, an improved version of the SURFNET program [131] and PSAIA (Protein Structure and Interaction Analyzer) [132].

DSSP is used for assigning secondary structure on a residue basis, given a protein structure. Secondary structure elements are recognized as repeats of the elementary hydrogen-bonding patterns "turn" and "bridge"; repeating turns are "helices", repeating bridges are "ladders" and connected ladders are "sheets".

DSSP is also used to provide the accessible surface area (ASA) of the protein under study, as the surface described by all possible positions of a water molecule (1.4Å radius) in contact with protein atoms, given on a residue basis as well. This follows the idea of Lee and Richards' water sphere rolling around the protein surface [133] and the algorithm subsequently developed by Shrake and Rupley [134].

Speedfill generates molecular surfaces of protein structures depicting cavities as convex regions, so that each atom cavity occurrence and average depth within a cavity can be determined. Interactions with molecules such as DNA, ligands and other proteins are mainly mediated by the shape and chemical properties of a protein surface [135] [136] [137] [138]; these characteristics also establish what will be the quaternary structure in multimeric proteins [139]. This program is particularly useful in molecular modeling and drug design [140] [141].

PSAIA is a software that makes the calculation of a protein geometry and the identification of protein-protein interaction sites easy, given a protein structure. PSA, the console version of PSAIA's structure analyzer module, is used to compute two geometrical parameters: The convexity or protrusion index (CX), and the depth index (DPX, which will be referred as burial index so that it is not confused with the depth of an atom or residue within a cavity).

The CX algorithm helps to identify protruding or convex atomic regions that are not only likely to be involved in protein-protein interactions, but can also help to identify limited proteolysis cleavage sites and antigenic determinants [142]. CX is computed as follows: A sphere of predetermined radius is centered around each heavy (non-hydrogen) atom; the number of other heavy atoms within the sphere is multiplied by the mean atomic volume found in proteins (20.1 ± 0.9 Å^3 [143]), which gives the volume occupied by protein within the sphere (internal volume); the free volume within the sphere (external volume, the difference between the sphere volume and the internal volume) is then divided by the internal volume, giving the CX; atoms in protruding regions have a high ratio (CX) between the external and the internal volume. In order to discern atoms in concave regions from those just buried, CX has to be combined with either an ASA calculation, or with DPX.

The DPX algorithm measures to what extent an atom is buried in the interior of a protein [144], something not covered by the ASA calculation alone. The burial of an atom is calculated by its distance in Å to the closest solvent accessible atom. Its values are higher than 0 for atoms buried in the protein core. These atoms buried in the protein core usually play a key role in the folding process, while contributing more to the protein thermodynamic stability; furthermore, buried atoms close to the surface, might become accessible through internal dynamics occurring upon binding, for example.

Structure coloring is achieved by another in-house Perl program named mappON, which implements the above mentioned software.

An independent version of mappON, the main output of which is a table with diverse parameters of the residues of interest, implements additional software, namely DisEMBL [145] to calculate protein disorder probability and retrieves evolutionary conservation and residue variability along evolution from the ConSurf-DB [146].

DisEMBL is an accurate method based on artificial neural networks that predicts potentially disordered/unstructured regions within a protein sequence. As no agreed definition of protein disorder exists, DisEMBL is trained for predicting the probability of disorder according to three definitions:

- Loops/coils as defined by DSSP; loops/coils are not necessarily disordered, albeit protein disorder is usually found within them. It follows that loop assignments are necessary but not sufficient for disordered segments.

- Hot loops, as a refined subset of the above, namely those with a high degree of mobility determined by Cα temperature factors.
- Missing coordinates in X-ray structures as defined by Remark465 entries in PDB; non assigned electron densities most often reflect intrinsic disorder.

ConSurf-DB is a repository of pre-calculated ConSurf [147] [148] [149] protein evolutionary conservation profiles for the whole PDB; evolutionary conservation and residue variability information is retrieved when available (otherwise, a close homolog for the query protein is searched for, in which case evolutionary conservation values are correlated to the query protein).

ConSurf maps the level of evolutionary conservation at each amino acid site based on the phylogenetic relations between the protein's or domain's close sequence homologues. The degree of conservation at each amino acid site is similar to the inverse of the site’s rate of evolution; slowly evolving sites are evolutionarily conserved, while rapidly evolving sites are variable. As previously commented, key amino acid positions that are important for maintaining the 3D structure of a protein (like residues buried in the protein core) and/or its function(s) (e.g. catalytic activity, binding to ligand, DNA or other proteins) are often under strong evolutionary constraints. Thus, the biological importance of a residue often correlates with its level of evolutionary conservation within the protein family.

This independent mappON version also calculates hydrogen bonding using DSSP and retrieves temperature factors from the structural files containing atomic coordinates. The distribution of temperature factors along a protein sequence is regarded as an important indicator of the protein's flexibility and dynamics. A large temperature factor indicates high mobility of individual atoms and side chains. Temperature factors have a variety of applications, such as predicting protein flexibility [150] [151], studying protein thermal stability [152] [153], analyzing active sites [154] [155] [156], correlating side-chain mobility with conformation [157] [158], analyzing protein disordered regions [159] [160] and investigating protein dynamics [161]. Temperature factors are computed as specified in [128]. They are taken from the structure for each atom in a residue, and then averaged over the whole residue. To exclude variations between proteins (measured temperature factors are given on different scales owing to the application of different refinement procedures [162]), the temperature factor of each residue is also standardized by subtracting the mean of all the temperature factors in the chain (except the highest and lowest value) and dividing the result by the standard deviation of all the temperature factors in the chain (except the highest and lowest value); standardized temperature factors are usually used in comparisons between different proteins and protein chains [150] [151] [163] [164]. Other atom-based parameters, such as average depth, are also averaged over the whole amino acid; in the case of cavity occurrences, since a residue might be part of more than one cavity, the cavity where a residue occurs is considered the largest where any of its atoms occur. Though the main output is a table with all these parameters of selected residues, mappON is also able to color structures according to these parameters.

This independent mappON version was applied to a phosphorylation site dataset in order to structurally characterize them and update the Phospho3D database, as detailed next.

2.6 - Phospho3D update and phosphorylation site predictor development

42,474 experimentally verified phosphorylation sites (in 8,718 eukaryotic proteins) stored in the Phospho.ELM database [165] (version 9.0, August 2010), both manually curated from the literature and obtained from mass spectrometry-based proteomics experiments were analyzed using mappON so as to derive the data to construct the new Phospho3D database [166] [167]. The latter is a repository of 3D structures of phosphorylation sites which stores information retrieved from Phospho.ELM and which is enriched with structural information and annotations at the residue level. The corresponding PDB structures of the proteins included in

Phospho.ELM were retrieved provided they fulfilled established thresholds (more than 98% sequence identity in non-gapped regions, more than 30 residues of the provided sequence correctly aligned, less than 15% of gaps present in the alignment, an e-value below 1e-6). Phosphorylation sites were mapped onto the corresponding structures; this resulted in 5,387 mapped instances (1,770 unique Phospho.ELM instances - 897 Serine, 338 Threonine, 535 Tyrosine - on 2,158 protein chains, mainly from human and mouse). Their structural context was analyzed according to the descriptors mentioned above; the same was done for the phosphorylation site neighboring residues (5 flanking residues on the left and 5 on the right, and those with their Cα within a sphere of 12Å radius).

The complete redundant list of all the PDB files fulfilling the thresholds was filtered using PISCES [123]. This list was filtered according to four increasing sequence identity thresholds: 30%, 50%, 90% and 100% and a large-scale structural analysis was performed using the latter, as mentioned in RESULTS.

Phospho3D stored phosphorylation sites' structural features as well as structure and sequence information of their neighborhood were used to train a structure-based phosphorylation site predictor based on a random forest algorithm. This was motivated by the fact that statistically significant differences were found between the structural descriptors of phosphorylation sites and control sites; control sites were defined as those residues sharing the same residue type (Serine, Threonine or Tyrosine) as the phosphorylation site in the same protein, but not annotated as phosphorylatable in Phospho.ELM.

The dataset used to build the predictor was balanced taking a random sample of control sites matching the number of mapped phosphorylation sites. Using the balanced data set, a fivefold cross-validation was performed; it involved partitioning the data set into two complementary subsets, one, comprising four fifths of the original data set, for setting the parameters (the training set), and the other, comprising the remaining fifth, for validating the analysis (the testing set). To reduce variability, five rounds of cross-validation were performed using different partitions, and the validation results were averaged over the rounds.

The analysis itself consisted in developing a random forest ensemble classifier, consisting of 5,000 decision trees that "vote" whether a residue instance is a phosphorylation site or not; the forest chooses the classification having the most votes (over all the trees in the forest). These trees are decision support tools using tree-like models that classify each instance considering different combinations of the following variables:

- Phosphorylation site's structural descriptors (namely secondary structure, solvent accessibility, temperature factor, cavity occurrence, average depth, CX, DPX, evolutionary conservation, residue variability and disorder according to 3 different criteria).

- Phosphorylation site's sequence flanking residues (5 before and 5 after).
- Structural descriptors for each of the 10 flanking residues.
- Frequencies of each of the 20 amino acids among the sequence flanking residues.
- Frequencies of each of the 20 amino acids among the structural residue neighbors within a 12Å radius sphere.

Seven different tests considering different combinations of the above variables were performed to finally find out that structural information did not provide an additional value to the descriptor, as discussed below. The predictor actually works at four different levels, considering either of the phosphorylatable residues Serine, Threonine or Tyrosine independently, or altogether. In all cases, structural information did not provide an additional value, as shown in RESULTS.

The predictor was developed using the R programming language for statistical computing and graphics.

2.7 - Database and web servers development

Web servers have been developed for the tools introduced so far: ModelDB, mappON and MAP. PHP, Javascript and Jmol script was applied in the former two cases, and HTML and Perl CGI (Common Gate Interface) in the latter. The former two web servers query a MySQL relational database which stores all the above mentioned information (basic information of each PDB structure, functional annotation and pre-calculated decoy sets and "colored" decoy sets, etc) and was built using the Perl DBI (DataBase Interface), which allows to embed database communication within Perl programs, and DBD (DataBase Driver) modules as plug-ins to DBI.

3 - RESULTS

3.1 - Database composition

The information stored in the ModelDB relational database comprises basic details of each PDB structure, such as title, macromolecule name, classification, biological source (organism), experimental determination method, resolution, chains, UniProt [168] [169] accession (chain-based, related to Swiss-Prot identifiers) and EC number (chain-based) if applicable; it also includes the functional information detailed above, coming from CREDO (which documents protein-ligand interactions derived from structures in the PDB), CSA (which stores information about enzyme active sites and catalytic residues, whenever an enzyme have a 3D structure available in the PDB as well) and Swiss-Prot. The database also stores zip files with pre-computed decoy sets (computed following the in-house ModelDB procedure), and "colored" decoy sets (modified using the in-house program mappON), as well as files that help to map relevant features onto native structures and decoy models (created using the in-house tool MAP). The simple organization is shown in Figure 5. This database can be obtained only upon request and built locally.

The complete database comprises the whole PDB as of 3rd January 2011; a 50% sequence identity culled subset was used to build decoy sets using a minimum of 80% coverage and a maximum e-value of 0.1 as HHsearch thresholds. 8,609 out of 179,636 PDB chains (in 68,442 PDB files) were modeled, but 1,442 cases yielded no model at all, as no template fullfilling the threshold was found. Only one case, 2CIO chain B, was discarded from the database. Out of the remaining 7,166 PDB chains, 2,999 have an EC number (72,648 in the whole PDB), 2,452 of them with a complete EC number of 4 digits (63,474 in the whole PDB); 5,106 bind to ligands (97,388 in the whole PDB), 3,742 of them to more than one (70,780 in the whole

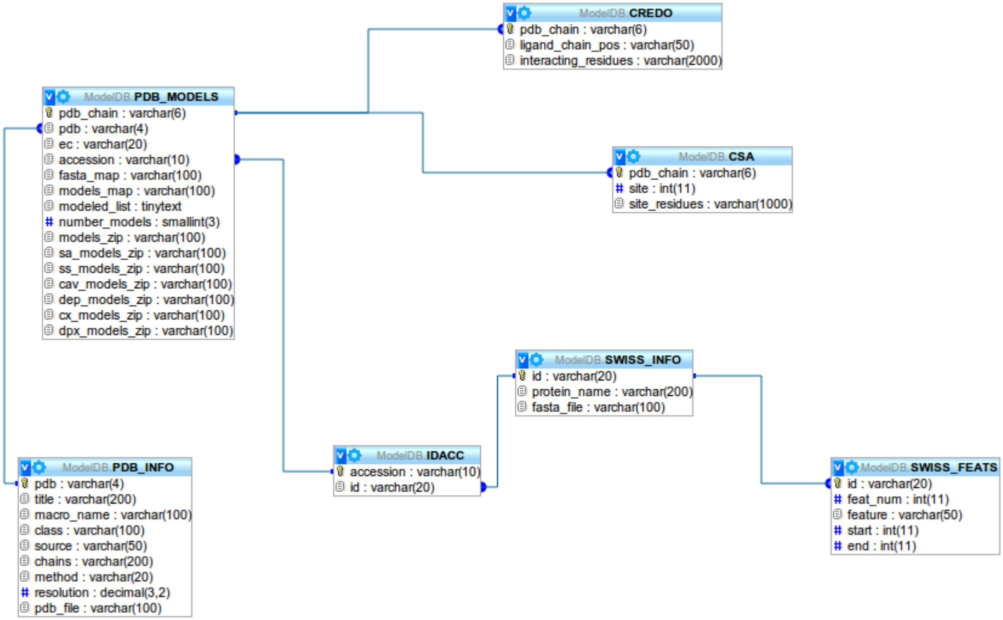


Figure 5: *Database design. Interestingly, the fasta_map and models_map fields in the table PDB_MODELS correlate native structure coordinates with Swiss-Prot coordinates and models coordinates, respectively; the "_zip" fields contain the zip files with models and colored models.*

PDB), and there are 1,199 different ligands found binding to this subset of modeled chains (9891 in the whole PDB); 2,261 have at least one catalytic site (51,437 in the whole PDB), 982 of them have more than one catalytic site (23,686 in the whole PDB); 4,546 are annotated in Swiss-Prot (125,966 in the whole PDB), 4,018 of them functional or biologically relevant annotation in Swiss-Prot (108,373 in the whole PDB), see METHODS section 2.4.

The average number of models per PDB chain in the ModelDB database is 17, being the highest number of models obtained 206 for 2RHE chain A; the number of decoy models computed for PDB chains follows the density distribution shown in Figure 6. It is worth considering, nonetheless, that as the PDB grows, new releases of ModelDB will include more models per decoy set. Plotting the GDT-TS mean and standard deviation for each PDB chain set of decoy models, one can get an idea of the GDT-TS range they span, as can be seen in Figure 7.

3.2 - ModelDB

ModelDB exists both as a local program for large calculations, and as a publicly available web server (see AVAILABILITY) with additional features.

3.2.1 - ModelDB program

The ModelDB pipeline was used to build the pre-calculated decoy sets stored in the database, out of a significant subset of the PDB. The user can provide either a single structure or a list of several structures; the sequence of each input structure (considered the native structure) is extracted and queried against a 70% non redundant PDB HMM database, implementing HHsearch. HHsearch hits are taken

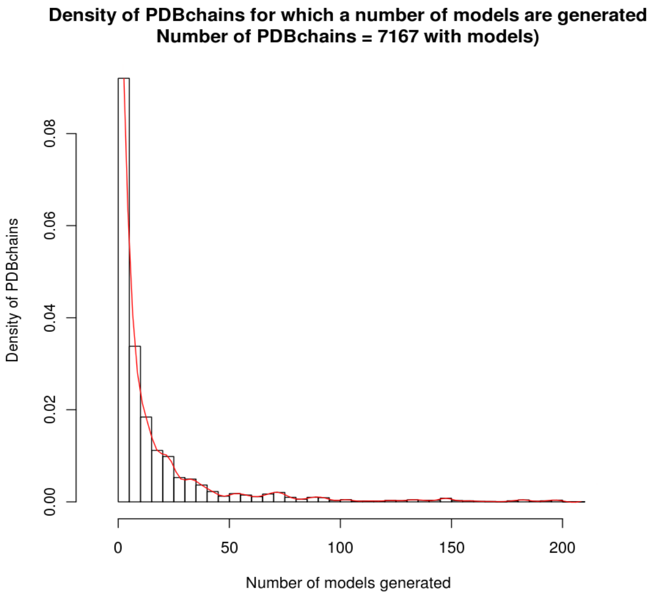


Figure 6: *Density distribution of the number of decoy models produced for 7,166 PDB chains in the input list. The majority of PDB chains have a number of decoy models between 0 and 10, being the average 17 and the maximum 206.*

as templates to build single-template models when fulfilling preselected thresholds; target-single template alignments are extracted from the HHsearch output and sent to Modeller to build all non-hydrogen atom single-template decoy models. Each decoy model is then compared to the native to assess its quality, in terms of GDT-TS and RMSD, applying LGA (Local-Global Alignment), useful for comparing two proteins or portions of them at the structural level. In order to be able to spatially superimpose all structures in a decoy set, every decoy is rotated and moved to best fit the native structure.

Thus, a resulting decoy set will comprise the native structure plus decoy models of different qualities that can be superimposed to it and ranked according to diverse quality estimators, either coming from LGA (GDT-TS, RMSD, LGA_S and LGA_Q with regard to the native structure), or from HHsearch (e-value, percentage identity and coverage with regard to the template used). Additionally, every structure in a decoy set can be "colored" according to one of the following six coloring schemes: Solvent accessibilities, secondary structures, cavity occurrences, average depths, protrusion indexes and burial indexes, using another in-house program named mappON.

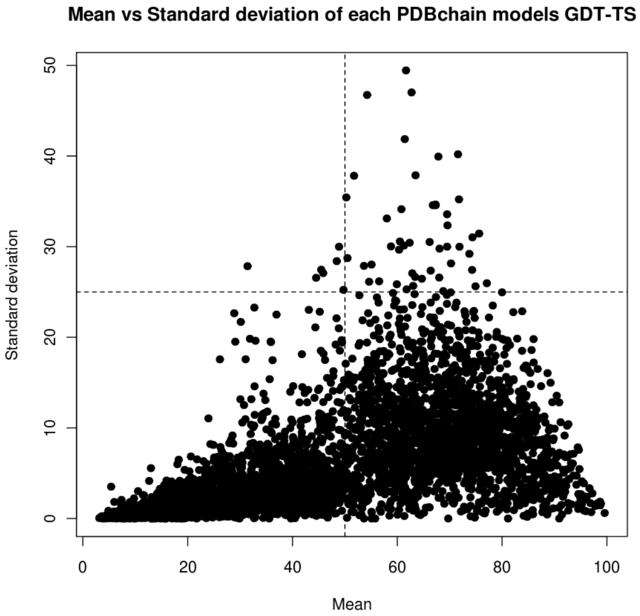


Figure 7: *Scatter plot representing the mean vs. the standard deviation of GDT-TS in each decoy set. The standard deviation gives an idea of the quality range in a decoy set.*

3.2.2 - ModelDB web server

The ModelDB web interface is conceived to be as user-friendly as possible. It has many additional features and benefits from the use of the database introduced above.

A user can either specify a PDB code or upload any protein structure of his/her own, specifying the chain of interest in both cases (otherwise the first chain present in the structure is analyzed); this structure is considered the native one in the set. Modeling default thresholds (80% coverage and an e-value of 0.1) can be changed. If a PDB code in the modeled list is selected without changing the default parameters, the user is redirected to the main page described next, where the pre-computed decoy models can be downloaded and visualized; however, if the PDB code is not part of the modeled list, the default parameters are changed or the structure under study is one uploaded by the user, there is an intermediate step where the modeling program is launched, and the models are computed. Once finished, the main page appears as well.

The ModelDB main page portrays in a comprehensible way the information stored in the database. An upper window gives a short description of the PDB code introduced (title, macromolecule name, classification, chains, crystallization method, resolution, EC number, UniProt accession, etc.); if the user introduced a structure of his own, a BLAST search is performed with stringent parameters (90% coverage and an e-value of 10^{-4}) against PDB and against Swiss-Prot (in case no PDB is found), so that basic information and functional annotation can be transfered from a close homolog to the structure under study.

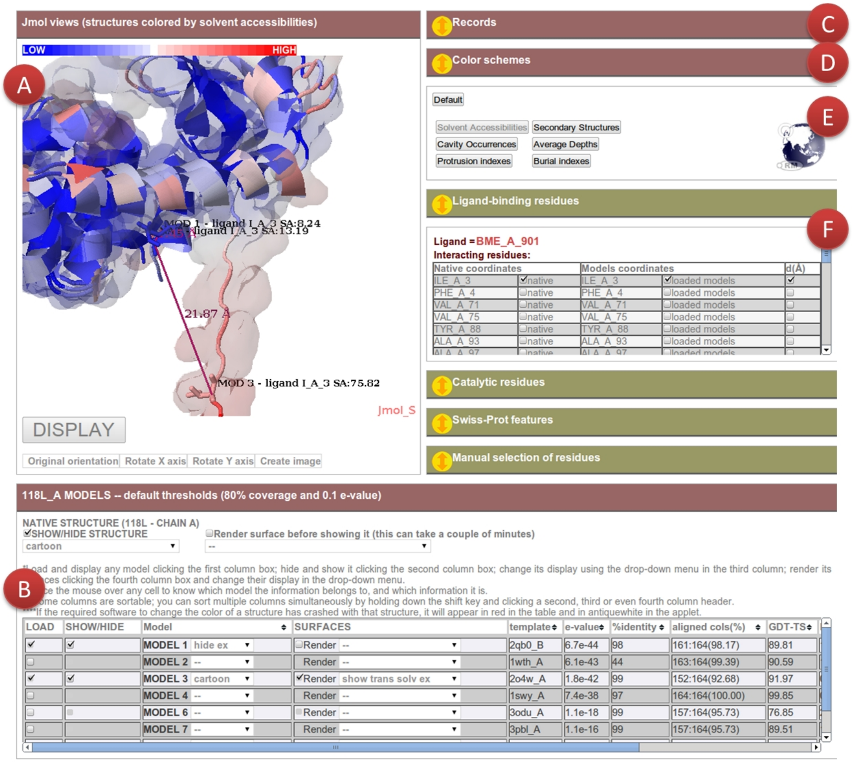
The possibility to download a zip file containing the decoy models produced is given, which include in their name the template used and diverse quality estimators for a straightforward ranking according to any of them. The main body of the page is composed by a sortable table (Figure 8.B) where the models in the decoy set are listed and can be ranked and visualized using a Jmol applet (Figure 8.A), where they can be loaded in any amount and order. Each model appears directly superimposed to the native structure. The native structure and the models are loaded in cartoons, but the display can be changed into spacefill, trace or backbone representations, etc.; besides, solvent excluded and solvent accessible surfaces can be rendered.

Collapsible boxes provide functional annotation as extracted from CREDO, CSA and Swiss-Prot (Figure 8.F); in the case of user uploaded structures, functional annotation is transfered from the PDB or Swiss-Prot entries of homologous proteins. Coordinate relations between native structure and models, between it and the corresponding Swiss-Prot fasta sequence (where applicable), and between it and the corresponding PDB structure (if it is a user uploaded structure with a PDB hit) are made using the MAP program. Biologically relevant residues plus any other that can be manually selected are highlighted as wireframe and labeled next to their Cα in the

Jmol window upon clicking the corresponding checkbox. The distance in Å between a native residue Cα and the Cα of the corresponding residue of each loaded model can be measured as well.

The page also features a state window which records everything happening in the Jmol applet (Figure 8.C), something useful for keeping track of the models loaded, displays, surfaces rendered, residues selected and distances measured. The user can also rotate the axes in the Jmol window and create images.

There is the possibility to color the structures and surfaces according to the six different coloring schemes (apart from the default where each structure has one different color) mentioned above (Figure 8.E). If the color scheme is changed, unless modified models are pre-computed, another intermediate window appears while the altered mappON software runs either DSSP, Speedfill or PSAIA. Once ready, the user is redirected to another main page analogous to the previous, where the applet is reloaded and the structures are colored according to the selected scheme (Figure 8). Here, once a residue is selected, the Cα value for the structural descriptor appears in its label and in the state window (except in the case of cavity occurrences, given on an atomic basis, where the largest cavity where any of the residue's atoms occurs is shown). Visualizing the models according to any of these color schemes helps understanding how these descriptors are distributed within a structure and amongst



Ligand =BME_A_901
Interacting residues:

Native coordinates		Models coordinates		d(Å)
ILE_A_3	☑native	ILE_A_3	☑loaded models	☑
PHE_A_4	□native	PHE_A_4	□loaded models	□
VAL_A_71	□native	VAL_A_71	□loaded models	□
VAL_A_75	□native	VAL_A_75	□loaded models	□
TYR_A_88	□native	TYR_A_88	□loaded models	□
ALA_A_93	□native	ALA_A_93	□loaded models	□
ALA_A_97	□native	ALA_A_97	□loaded models	□

118L_A MODELS -- default thresholds (80% coverage and 0.1 e-value)

NATIVE STRUCTURE (118L - CHAIN A)
☑SHOW/HIDE STRUCTURE
cartoon ▾
□Render surface before showing it (this can take a couple of minutes)
-- ▾

*Load and display any model clicking the first column box; hide and show it clicking the second column box; change its display using the drop-down menu in the third column; render its ces clicking the fourth column box and change their display in the drop-down menu.
ce the mouse over any cell to know which model the information belongs to, and which information it is.
ome columns are sortable; you can sort multiple columns simultaneously by holding down the shift key and clicking a second, third or even fourth column header.
****If the required software to change the color of a structure has crashed with that structure, it will appear in red in the table and in antiquewhite in the applet.

LOAD	SHOW/HIDE	Model		SURFACES		template	e-value	%identity	aligned cols(%)	GDT-TS
☑	☑	MODEL 1	hide ex ▾	□Render	-- ▾	2qb0_B	6.7e-44	98	161:164(98.17)	89.81
□		MODEL 2	-- ▾	Render	-- ▾	1wth_A	6.1e-43	44	163:164(99.39)	90.59
☑	☑	MODEL 3	cartoon ▾	☑Render	show trans solv ex ▾	2o4w_A	1.8e-42	99	152:164(92.68)	91.97
□		MODEL 4	-- ▾	Render	-- ▾	1swy_A	7.4e-38	97	164:164(100.00)	99.85
□	□	MODEL 6	-- ▾	□Render	-- ▾	3odu_A	1.1e-18	99	157:164(95.73)	76.85
□		MODEL 7	-- ▾	Render	-- ▾	3pbl_A	1.1e-16	99	157:164(95.73)	89.51

Figure 8: *Section of the ModelDB result page described in the text. The structure of 118l chain A (T4 lysozyme) and its 3rd model are displayed in the Jmol window as cartoons and colored by solvent accessibilities, the 3rd model's transparent solvent excluded surface is also depicted; a ligand binding Isoleucine is highlighted in both structures, as well as in the 1st model; 3rd model's Isoleucine occurs in an unfolded loop, far away from where it should be, making it more accessible.*

structures of different quality.

The ModelDB package, available for download, includes, apart from the mappON and MAP programs, some other in-house programs that can be used independently as well: A program that extracts sequences from structures and runs BLAST against PDB and/or Swiss-Prot; a program that, instead of building a decoy set out of a protein structure, builds a more refined single model out of a protein sequence (it builds an HMM as query and uses more precise parameters for Modeller); and another program that benefits from the use of LGA to rotate and move one structure in order to fit another one, something very useful for creating pictures using a molecular visualization program like Jmol or PyMol.

3.2.3 - Examples of use

The decoy sets are conceived to test structure-based methods and define to which extent they can make use of predicted protein structure models. However, the functional documentation, the model quality estimates and the different color schemes allow many large-scale analyses to be performed as well; to serve as example, I checked until which level of model accuracy I could still detect the same exposed and buried residues, the same residues defining the largest cavity, etc. as in the native structure, as explained next.

3.2.3.1 - Exposed and buried residues detection

Using all the decoy sets produced for the PDB subset, I studied to what extent exposed residues are still detectable across models of decreasing quality. Exposed residues are considered those with a normalized solvent accessibility value above 70%; solvent accessibilities are normalized according to maximum residue values, as defined by Miller *et al*. [170]). More than 75% of exposed residues are still detected in over 40% of models with a GDT-TS above 90, and in almost 30% of those with a GDT-TS above 80; beyond this value of GDT-TS, there is a drop of the percentage of models where at least 75% of the exposed residues are detectable, barely reaching 10% (Figure 9).

Buried residues, considered in this analysis those with a normalized solvent accessibility below 30%, tend to be more conserved along evolution than exposed ones, as they usually play a key role in the folding process, while contributing more to the protein thermodynamic stability. Thus, models build by homology often maintain these residues buried, no matter their overall quality (Figure 9).

3.2.3.2 - Largest cavity detection

I have also focused on the most important subset of exposed residues, those forming the largest protein cavity. As studied for some examples by Kuntz *et al.* [171], binding sites tend to occur in the largest cavity; hence, this region defines the

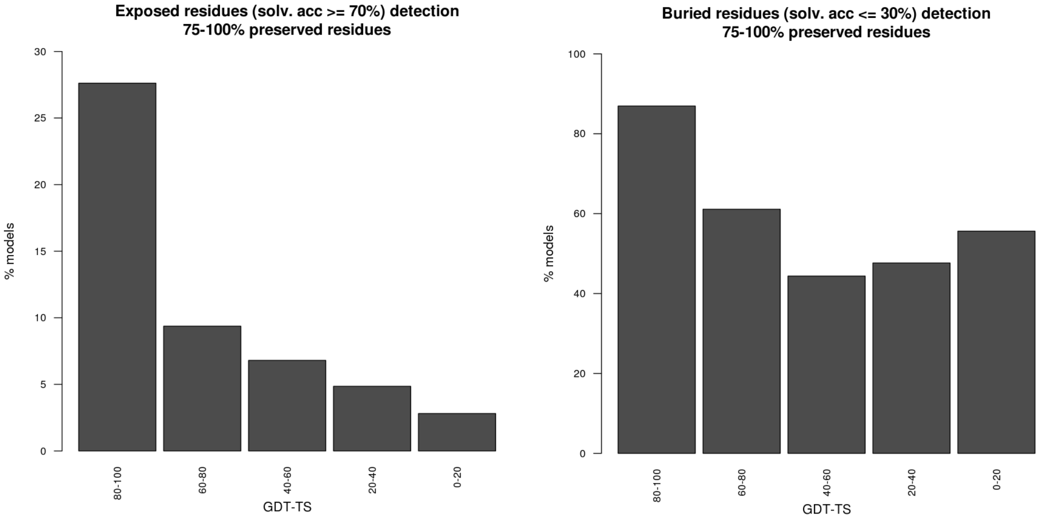


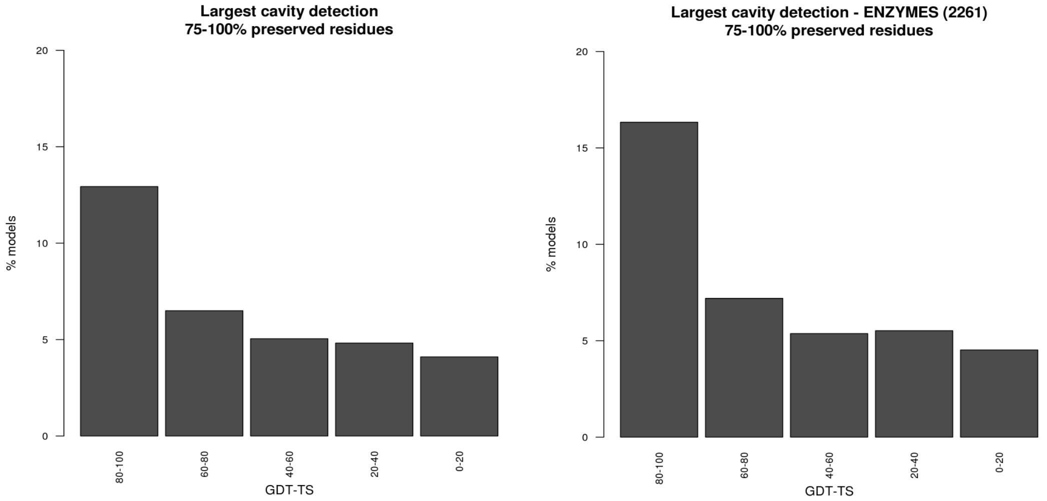
Figure 9: *Exposed (left) and buried (right) residues detection as model quality decreases, in terms of GDT-TS; residues in the protein core are better maintained in their buried positions.*

space available for docking new ligands or modifying existing ones, something of paramount importance in the fields of molecular modeling and drug design. The largest cavity is slightly less detectable than the whole set of exposed residues; only in a 20% of the models with a GDT-TS above 90 can at least 75% of the residues constituting such cavity be detected (a residue is considered to belong to a cavity if any of its atoms belong to it); there is a significant drop when GDT-TS is lower than 80 (Figure 10). However, such largest cavity is better detected in the subset of enzymes stored in the CSA (Figure 10).

We have studied whether catalytic residues occurring in the largest cavity of enzymes can still be detected as model quality drops; since the precise position of these residues tends to be well maintained in an enzyme family, models built by homology have such residues still constituting part of the largest cavity, no matter the model accuracy (Figure 11).

3.2.3.3 - Active site residue relative position

Another study carried out consisted in measuring the Euclidean distance



Largest cavity detection
75-100% preserved residues

Largest cavity detection - ENZYMES (2261)
75-100% preserved residues

Figure 10: *Largest cavity detection in the decoy models of the whole PDB subset (left) and in those considered enzymes in the CSA (right), where the cavity is slightly better detected.*

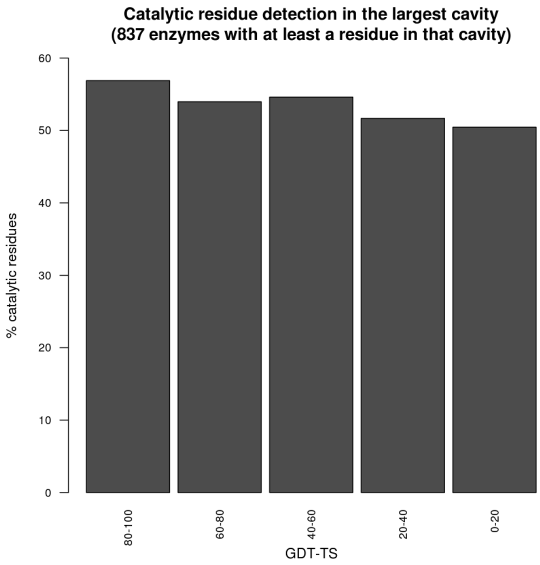


Figure 11: *When a catalytic residue occurs in the largest cavity, it is well maintained there in an enzyme family; thus, decoy models built by homology, will maintain those catalytic residues in the largest cavities as well.*

differences between every permutation of catalytic residue Cαs constituting an active site (with two or more residues) in the native structure and the same permutation in its models of varying quality. Averaging all the differences per permutation over the active site as a whole showed an increasing mean Euclidean distance difference as model quality decreases in terms of GDT-TS, meaning that modeled catalytic sites are increasingly further apart from their native counterpart (Figure 12). Nonetheless, the maximum mean value per site (when model quality is the lowest), is slightly higher than 0.5Å, meaning that catalytic residues relative positions tend to be maintained even in low quality models (Figure 12).

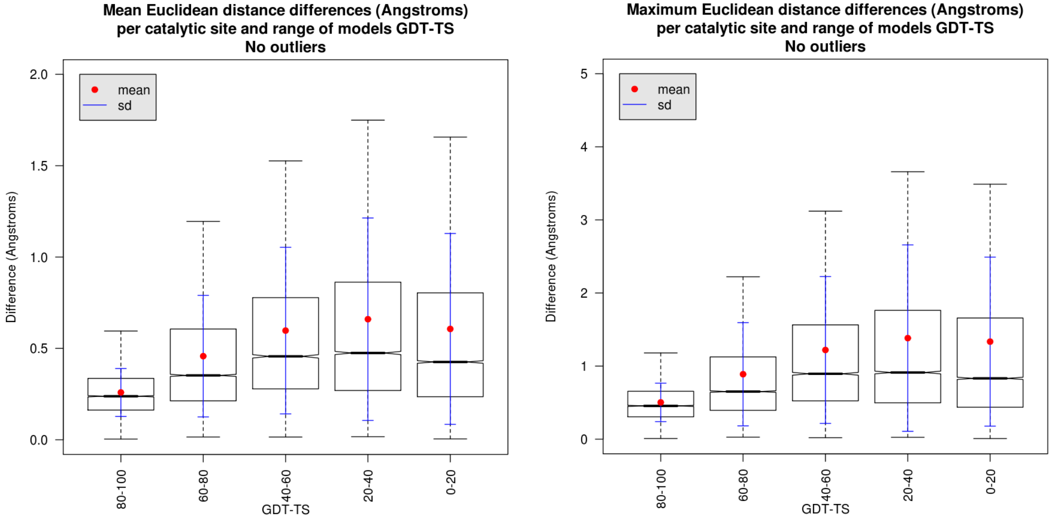


Figure 12: *Mean (left) and maximum (right) Euclidean distance differences per catalytic site in CSA and range of models GDT-TS. Catalytic sites tend to be slightly further apart from native counterparts as model quality drops, but not so far apart, being the relative positions maintained.*

3.3 - MAP program and web server

The bl2seq-based MAP local program and public web server (see AVAILABILITY) have some features intended to deal with the common problem in bioinformatics of mapping sequence residues onto structures, or structure residues onto another structure. It is part of both the ModelDB and mappON web servers. The MAP web server itself is specially conceived to be the most user-friendly possible; it offers 5 possibilities:

1. Extract ATOM sequence in fasta format: Upon uploading a PDB protein file and specifying a chain of interest (otherwise the first one present in the file is taken), the atomic coordinate sequence is taken and the amino acid sequence is extracted in fasta format.
2. Number ATOM residues: The atomic coordinate sequence is extracted from a structure and the residues are renumbered consecutively starting from 1.
3. Map coordinates from fasta sequence to structure: This is the main reason why this program/web server was developed, and the feature applied in both ModelDB and mappON. Upon uploading a fasta sequence and a corresponding structure, the residues of the former are mapped onto the coordinates of the latter. If only one sequence is uploaded, the corresponding structure is retrieved via a BLAST search with stringent parameters (90% coverage and an e-value of 10^{-4}). The output consists of a table of correspondences between sequence and structure residues. Mapping residues from sequence to structure is far from trivial, as residues might not be visible in the structure.
4. Map coordinates from one fasta sequence to another.
5. Map coordinates from one structure to another: Upon providing two

structures (or just one, and retrieving the other using BLAST), the user gets a table of residue correspondences between the structures.

3.4 - MappON

The mappON tool is implemented in the ModelDB pipeline to color every decoy in a set. Another version of mappON exists independently both as a local tool for large calculations and as a publicly available web server (see AVAILABILITY), and offers some other features. Apart from coloring input structures according to diverse descriptors, it outputs a table with the descriptors of selected residues (and those surrounding them). Thus, it serves to analyze properties of key residues in the protein structural context and visually examine the results.

3.4.1 - MappON program

MappON integrates several algorithms selected for their efficient performance or their uniqueness in their kind, making the use of them straightforward and fast, and providing access to some not readily available.

The original mappON version takes as input either a structure or a sequence, or both; PDB structures are searched using BLAST with stringent parameters (90% coverage and an e-value of 10^{-4}) whenever a sequence alone is used as input. Given the location of relevant residues based on the sequence or on the structure (when it is provided alone), mappON returns a file that provides data on each selected residue in terms of secondary structure, solvent accessibility, cavity occurrence, average depth, protrusion index and burial index, hydrogen bonding, temperature factor, evolutionary conservation, residue variability and disorder probability.

MappON also allows the study of the residue neighborhood of any selected residue. This neighborhood comprises any number of flanking residues in sequence, plus those with their Cα within a sphere of any radius centered on the Cα of the selected residue. An additional program in the mappON package allows coloring input structures following color schemes defined by any of the possible above listed descriptors.

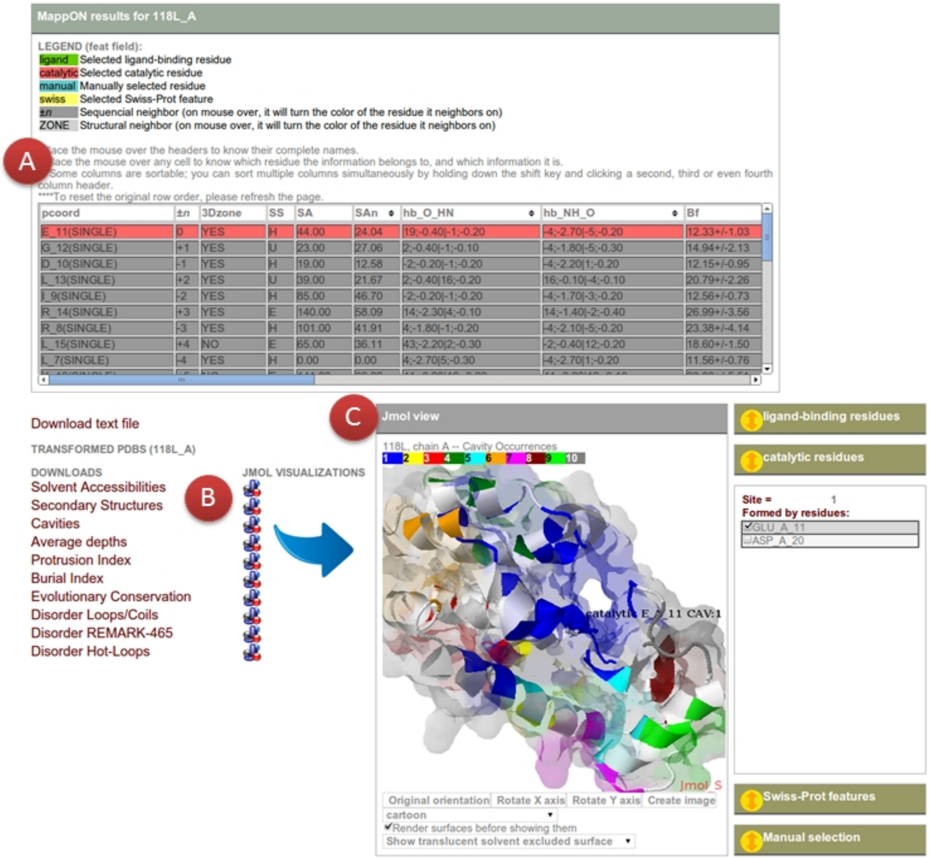
In practice mappON represents a tool that integrates several structural and functional information that, in turn, is useful to investigate properties of the protein under study and of its residues.

3.4.2 - MappON web server

The mappON web server is as user-friendly as possible and has an interface very similar to that of ModelDB, providing the same functional annotation stored in the database, and making use of the MAP program as well. The server gives the possibility to either specify a PDB code, upload a structure, or a fasta sequence. Upon uploading a file, a BLAST search against PDB (and against Swiss-Prot if the file is a structure and there are no hits in PDB) is performed so as to be able to provide functional annotations (mapped using the MAP program).

A primary page allows the selection of biologically relevant residues (ligand-binding, catalytic and others, as previously specified) and gives the possibility to study a residue neighborhood for each selected instance, comprising any number of residues in sequence and those within a specified radius sphere.

The result page contains a sortable table listing the multiple residue-based structural/functional descriptors mentioned above, for each of the selected residues and neighboring residues (Figure 13.A). Apart from the result table, the user obtains the protein structure modified so that its temperature factor fields are exchanged with each of the calculated descriptors, useful to visually examine residue characteristics via a molecular viewer, such as PyMol or Jmol. These modified structures (in cartoons) with the selected residues highlighted (in wireframe with labels indicating their descriptor value) can be popped-up in the result page (Figure 13.B) in a Jmol window, allowing the highlighting of any other of the biologically relevant residues. The popped-up structure display can be changed, and solvent excluded and solvent accessible surfaces can be rendered (Figure 13.C).



MappON results for 118L_A

LEGEND (feat field):

- 🟩 ligand Selected ligand-binding residue
- 🟥 catalytic Selected catalytic residue
- manual Manually selected residue
- swiss Selected Swiss-Prot feature
- ±*n* Sequencial neighbor (on mouse over, it will turn the color of the residue it neighbors on)
- ZONE Structural neighbor (on mouse over, it will turn the color of the residue it neighbors on)

lace the mouse over the headers to know their complete names.
lace the mouse over any cell to know which residue the information belongs to, and which information it is.
Some columns are sortable; you can sort multiple columns simultaneously by holding down the shift key and clicking a second, third or even fourth column header.
****To reset the original row order, please refresh the page.

pcoord	±*n*	3Dzone	SS	SA	SAn	hb_O_HN	hb_NH_O	Bf
E_11(SINGLE)	0	YES	H	44.00	24.04	19;-0.40|-1;-0.20	-4;-2.70|-5;-0.20	12.33+/-1.03
G_12(SINGLE)	+1	YES	U	23.00	27.06	2;-0.40|-1;-0.10	-4;-1.80|-5;-0.30	14.94+/-2.13
D_10(SINGLE)	-1	YES	H	19.00	12.58	-2;-0.20|-1;-0.20	-4;-2.20|1;-0.20	12.15+/-0.95
L_13(SINGLE)	+2	YES	U	39.00	21.67	2;-0.40|16;-0.20	16;-0.10|-4;-0.10	20.79+/-2.26
I_9(SINGLE)	-2	YES	H	85.00	46.70	-2;-0.20|-1;-0.20	-4;-1.70|-3;-0.20	12.56+/-0.73
R_14(SINGLE)	+3	YES	E	140.00	58.09	14;-2.30|4;-0.10	14;-1.40|-2;-0.40	26.99+/-3.56
R_8(SINGLE)	-3	YES	H	101.00	41.91	4;-1.80|-1;-0.20	-4;-2.10|-5;-0.20	23.38+/-4.14
L_15(SINGLE)	+4	NO	E	65.00	36.11	43;-2.20|2;-0.30	-2;-0.40|12;-0.20	18.60+/-1.50
L_7(SINGLE)	-4	YES	H	0.00	0.00	4;-2.70|5;-0.30	-4;-2.70|1;-0.20	11.56+/-0.76

Download text file

TRANSFORMED PDBS (118L_A)

DOWNLOADS

- Solvent Accessibilities
- Secondary Structures
- Cavities
- Average depths
- Protrusion Index
- Burial Index
- Evolutionary Conservation
- Disorder Loops/Coils
- Disorder REMARK-465
- Disorder Hot-Loops

JMOL VISUALIZATIONS

Jmol view

118L, chain A -- Cavity Occurrences

Site = 1
Formed by residues:
☑ GLU_A_11
□ ASP_A_20

Figure 13: *Section of the mappON result page described in the text. The catalytic Glutamic Acid, as well as its sequence and structural neighboring residues, of the structure 118L chain A (T4 lysozyme) has been analyzed. The structure of the lysozyme is shown colored by cavity occurrences; the catalytic Glutamic Acid belongs to the largest cavity.*

3.4.3 - Phospho3D 2.0

The mappON program has already proved its usefulness as it contributed to update the Phospho3D database [167], which stores phosphorylation sites in available protein 3D structures, along with information retrieved from the Phospho.ELM database [165] and descriptors obtained using mappON.

The original Phospho3D database [166] was already enriched with structural annotation at the residue level, including ASA, secondary structure and residue conservation extracted from the Consurf-HSSP database [172]. This 1.0 version also collected the annotation of the phosphorylayion site flanking sequence (10 residues) and a 12Å radius phospho-instance 3D neighborhood in structural space. It is known that these neighboring residues contact the kinase active site, affect the phosphorylation mechanism and regulate/modulate the specificity of the kinase interaction, but it is not clear whether the sequence neighbours are more relevant than the structural ones.

Since then, more than 26,000 structures have been deposited in the PDB and the number of Phospho.ELM instances has increased about fourfold. Phospho3D version 2.0 includes an eleven fold increase in the number of Phospho.ELM unique instances mapped onto 3D structures (compared to version 1.0), and several novel features, including all the structural descriptors of phosphorylation sites (as well as of the flanking residues at sequence level and those within a 12Å radius sphere) derived using the mappON tool, the possibility of browsing the database selecting non-redundant sets of 3D structures, the availability of downloading many non-redundant sets of structurally annotated phosphorylation sites (meant to serve as reliable benchmark datasets for phosphorylation site predictors' training and testing) and P3Dscan, a new functionality that allows the user to submit a protein structure and scan it against the 3D phosphorylation site zones collected in the Phospho3D database.

3.4.4 - Phosphorylation sites structural characterization

A large-scale structural analysis on the phosphorylation sites stored in the new Phospho3D database and mapped onto non-identical PDB structures was performed using mappON. The different plots obtained can be found at http://www.phospho3d.org/stats.py#3. The majority of the Phospho.ELM instances are in proteins of unknown structure, more than one quarter are in sequences belonging to proteins of known structures, but in regions for which coordinates are not available; only a small percentage can be reliably mapped onto a 3D structure. There are 1,770 unique Phospho.ELM sites mapped 5,387 times onto 2,158 different protein chains; phosphorylation sites mapped more than once on different structures were considered different phosphorylation sites in the analysis. The analysis was carried out separately for each set, plotting the statistical distribution of each 3D attribute used to annotate the phosphorylation sites in the database; here, only the plots for the phosphorylation sites falling on non-identical structures are reported.

Only a few instances are mapped on structure residues that are in the phosphorylated state in the structure (i.e. showing the attached phosphate group). A fraction of P-sites shows low accessibility to the solvent (in the interval 0-20%); however, the majority of them has accessibility higher than 20%. The majority of phosphorylation sites occur in loops or unassigned regions; the latter mean non determined secondary structure as defined by the DSSP program, and they are interpreted as loop or irregular elements. Two of the three disorder criteria (the hot-loops criterion is more stringent) agree on classifying many of the phosphorylation sites as disordered or flexible residues. They are not protruding and not buried within the protein core and, when located within a cavity, this is often the largest one. Besides, phosphorylation sites tend to be subject to strong evolutionary constraints.

Being phosphorylation a modification that involves protein-protein interaction, these results are consistent with the fact that phosphorylation sites should be found in a kinase recognition cleft, surrounded by flexible regions able to fit the kinase active site, and accessible on the surface of proteins, not buried within globular domains, as they have to be directly contacted.

3.4.5 - Phosphorylation site predictor

A random forest-based phosphorylation site predictor benefiting from both sequence and structural information was built using Phospho3D data obtained with mappON. The development of this predictor was motivated by significant structural differences found between phosphorylation sites and control sites (those residues sharing the same residue type as a phosphorylation site in the same protein, but not annotated as phosphorylatable in Phospho.ELM). Phosphorylation sites were found to have significantly higher solvent accessibilities, temperature factors, average depths and disorder, lower burial indexes, were found more often in loop regions and larger cavities, and more evolutionary conserved (data not shown); this is consistent with the structural features expected for a phosphorylation site, given the fact that phosphorylation involves protein-protein interaction.

Different combinations of variables were considered in the predictive analysis, including structural features of phosphorylation sites and neighboring residues in sequence and within a 12Å radius sphere, as well as sequence information of the sequence neighborhood and frequencies of amino acids in the sequence and structural neighborhood.

Other phosphorylation site predictors using signature 3D profiles, in particular, recently developed programs Phos3D [173] and PHOSIDA (Phosphorylation Site Database) [174] [175], claim to be more accurate than predictors using sequence information only. Both of these predictors have proved the added value (small but consistent) of using spatial information for the computational prediction of phosphorylation sites applying Support Vector Machines (SVMs) with both sequence and structural information, outperforming many other predictors, not only sequence-only based. SVMs are an ensemble of algorithms that analyze data and recognize patterns, used for classification and regression studies; the standard SVM takes a set of input data and predicts, for each given input, to which of two possible classes the input belongs. In other words, a SVM is a non-probabilistic binary linear classifier.

My random forest-based predictor can compete with Phos3D and PHOSIDA in terms of accuracy. PHOSIDA has an accuracy of 90,17% for the prediction of Serine sites (89,85% using sequence information alone) and an accuracy of 77,27% for Threonine sites (74,24% using sequence information alone). Phos3D has kinase-specific prediction accuracies ranging from 0,69 (Thr kinases) to 0,89 (MAP kinases). My predictor reaches an overall accuracy of 0,795 (residue type independent prediction), an accuracy of 0,786 for Serine sites, 0,829 for Threonine sites and 0,832 for Tyrosine sites. However, accuracies were actually higher when using sequence information alone (0,855 for residue type independent prediction, 0,810 for Serine sites, 0,833 for Threonine sites and 0,884 for Tyrosine sites). As anticipated in the introduction, it is thought that phosphorylation sites are an example of biologically relevant sites which sequence neighborhood has been subject to stringent evolutionary constraints; kinase recognition determinants correspond to consensus sequences with conserved residues that play a key role in the process.

4 - CONCLUSIONS

Since the gap between known protein sequences and structures continues to increase, researchers need to make use of protein structural models more routinely. Models usually contain structural inaccuracies that vary in number and severity, but they should not be discarded, as they can still provide important insights into a protein role. Any robust structure-based method able to effectively use these suboptimal models, namely those in the midrange and low range of quality, will be at an advantageous position and will have an inestimable practical value. Thus, there is the need to test the model quality tolerance for such methods. ModelDB, the tool introduced here, serves this purpose by generating decoy sets in a straightforward fashion. These decoys, or computed protein structure models of different qualities, are intended to be used to test structure-based methods and decide to which extent those methods can be applied to computed protein structure models. In other words, these decoys can be used to benchmark the applicability of a given method to models, deriving a quality threshold at which interpretable results analogous to the ones that would be obtained with native structures can be produced.

The project has involved the implementation of a pipeline divided in programs that work together, but also exist independently, either on-line or for local use when larger calculations are demanded. The ModelDB modeling pipeline takes a protein structure as input to generate single-template decoy models and rotate and move them to best fit the input native structure; it makes use of another in-house program named mappON to "color" the structures according to different spatial descriptors (solvent accessibilities, cavity occurrences, etc.). The on-line versions of both ModelDB and mappON query a relational database that not only contains pre-calculated decoy models and colored decoy models, but also functional annotations extracted from different sources and related to every structure in a decoy set using another in-house program named MAP.

The independent version of mappON is useful for the structural analysis of given residues as well as their residue neighborhood in sequence and space, listed in a tabulated output; the on-line version, in particular, provides annotation for biologically relevant residues and features a visualization window where colored structures with selected residues highlighted can be manipulated and visually validated. MAP is a basic tool that strives to deal with the common problem in bioinformatics of correlating residues from a protein sequence to a structure and between structures, something particularly easy in the on-line version.

There are no other publicly available resources for generating decoy sets in such an easy and user-friendly way, to our record; some other sites exits for just retrieving decoy sets, the most used ones being Decoys 'R' Us or the CASP page where the models submitted for each target can be downloaded. ModelDB allows the computation of new decoy sets of certain proteins a researcher might be interested in. Besides, since decoy models have been already created for a significant subset of the PDB, ModelDB covers a larger portion of the protein structural space compared to the other resources; this portion increases as new decoy sets are built and stored in the database. Individual decoy sets themselves are expected to cover wider quality ranges in new releases as more structures are deposited in the PDB. Last but not least, ModelDB also features a visualization window where any decoy in a set, colored according to different descriptors, can be loaded, inspected and compared with the native counterpart.

Apart from ModelDB's main application (i.e. benchmarking the applicability of computed models), the functional documentation, the model quality estimates and the structures color schemes allow several large-scale analyses, as shown in RESULTS using different examples.

5 - AVAILABILITY

Web servers introduced are publicly available at the following addresses:

1. ModelDB:
   http://bl210.caspur.it/MODEL-DB/MODEL-DB_web/MODindex.php
2. mappON:
   http://bl210.caspur.it/MODEL-DB/mappON_web/mappONindex.php
3. MAP:
   http://bl210.caspur.it/MODEL-DB/MAP_web/paginasDANIEL/MAP.html

Corresponding programs for local use can be downloaded in these web servers.

6 - ACKNOWLEDGEMENTS

I am thankful to Anna Tramontano for giving me the opportunity to join her Biocomputing Group at the Sapienza University of Rome, the Pasteurian Sciences School for useful seminars and discussions and the King Abdullah University of Science and Technology for the funding.

Agradezco especialmente el incondicional apoyo de mis padres, Mª Ángeles y Fernando, mi hermano Pablo, Carmen, Tita e Ión, a quienes va dedicado este trabajo y todo el esfuerzo en estos años de plomo. Al resto de mi familia, estén con nosotros o no (abuelos no os olvido); a todos mis amigos y compañeros de aventuras en Madrid, ya sean de mi barrio o de mi universidad (sobre todo a los que habéis venido a verme), y a Ahmed, que de tan pesado le he cogido cariño. Carmen, yo tampoco olvidaré los paseos por Roma y nuestro pequeño palacio.

Siempre viene bien acabar un trabajo con una cita; siempre hay alguien que lo ha hecho mejor que tú y si no puedes superarlo, róbaselo y aprovéchate; así que ahí va, esto resume muy bien estos años (por suerte tres y no cien):

> Había estado en la muerte, en efecto, pero había regresado porque no pudo soportar la soledad. Repudiado por su tribu, desprovisto de toda facultad sobrenatural como castigo por su fidelidad a la vida, decidió refugiarse en aquel rincón del mundo todavía no descubierto por la muerte, dedicado a la explotación de un laboratorio de daguerrotipia.

Gabriel García Márquez, 1967 - Cien años de soledad.

7 - BIBLIOGRAPHY

[1] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112(3): 535-42.

[2] Bairoch A, Apweiler R (1996). The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res* 24(1): 21-5.

[3] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223): 496-512.

[4] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860-921.

[5] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al. (2001). The sequence of the human genome. *Science* 291(5507): 1304-51.

[6] Lesk AM, Chothia C (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 136(3): 225-70.

[7] Chothia C, Lesk AM (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J* 5(4): 823-6.

[8] Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci U S A* 104(29): 11963-8.

[9] Kosloff M, Kolodny R (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins* 71(2): 891-902.

[10] Hou J, Jun SR, Zhang C, Kim SH (2005). Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc Natl Acad Sci U S A* 102(10): 3651-6.

[11] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997). CATH--a hierarchic classification of protein domain structures. *Structure* 5(8): 1093-108.

[12] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic*

Acids Res 25(17): 3389-402.

[13] Brenner SE, Chothia C, Hubbard TJ (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A* 95(11): 6073-8.

[14] Chen Z (2003). Assessing sequence comparison methods with the average precision criterion. *Bioinformatics* 19(18): 2456-60.

[15] Reid AJ, Yeats C, Orengo CA (2007). Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. *Bioinformatics* 23(18): 2353-60.

[16] Ohlson T, Wallner B, Elofsson A (2004). Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins* 57(1): 188-97.

[17] Eddy SR (1998). Profile hidden Markov models. *Bioinformatics* 14(9): 755-63.

[18] Söding J (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7): 951-60.

[19] Jaroszewski L, Rychlewski L, Godzik A (2000). Improving the quality of twilight-zone alignments. *Protein Sci* 9(8): 1487-96.

[20] Söding J, Biegert A, Lupas AN (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33(Web Server issue): W244-8.

[21] Sadreyev RI, Baker D, Grishin NV (2003). Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Sci* 12(10): 2262-72.

[22] Dunbrack RL Jr (2006). Sequence comparison and protein structure prediction. *Curr Opin Struct Biol* 16(3): 374-84.

[23] Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM (2003). A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins* 53(Suppl 6): 369-79.

[24] Kosinski J, Gajda MJ, Cymerman IA, Kurowski MA, Pawlowski M, Boniecki M, Obarska A, Papaj G, Sroczynska-Obuchowicz P, Tkaczuk KL, Sniezynska P, Sasin JM, Augustyn A, Bujnicki JM, Feder M (2005). FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. *Proteins* 61(Suppl 7): 106-13.

[25] Koliński A, Bujnicki JM (2005). Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins* 61(Suppl 7): 84-90.

[26] Zhang Y (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* 69(Suppl 8): 108-17.

[27] Zhang Y, Skolnick J (2005). The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci U S A* 102(4): 1029-34.

[28] Chothia C, Lesk AM. (1982). Evolution of proteins formed by beta-sheets. I. Plastocyanin and azurin.. *J Mol Biol* 160(2): 309-23.

[29] Read RJ, Brayer GD, Jurásek L, James MN (1984). Critical evaluation of comparative model building of Streptomyces griseus trypsin. *Biochemistry* 23(26): 6570-5.

[30] Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29: 291-325.

[31] Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A (2006). Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* Chapter 5(Unit 5.6).

[32] Sali A, Blundell TL (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3): 779-815.

[33] Fiser A, Do RK, Sali A (2000). Modeling of loops in protein structures. *Protein Sci* 9(9): 1753-73.

[34] Tress M, Ezkurdia I, Graña O, López G, Valencia A (2005). Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* 61(Suppl 7): 27-45.

[35] Baker D, Sali A (2001). Protein structure prediction and structural genomics. *Science* 294(5540): 93-6.

[36] Wallner B, Elofsson A (2006). Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci* 15(4): 900-13.

[37] Wallner B, Larsson P, Elofsson A (2007). Pcons.net: protein structure prediction meta server. *Nucleic Acids Res* 35(Web Server issue): W369-74.

[38] Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003). 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19(8): 1015-8.

[39] Sánchez R, Sali A (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A* 95(23): 13597-602.

[40] Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis F, Rossi A, Marti-Renom MA, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A (2006). MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* 34(Database issue): D291-5.

[41] Chen H, Kihara D (2008). Estimating quality of template-based protein models by alignment stability. *Proteins* 71(3): 1255-74.

[42] Tress ML, Jones D, Valencia A (2003). Predicting reliable regions in protein alignments from sequence profiles. *J Mol Biol* 330(4): 705-18.

[43] Kalman M, Ben-Tal N (2010). Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics* 26(10): 1299-307.

[44] Sippl MJ (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des* 7(4): 473-501.

[45] Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7: 95-9.

[46] Melo F, Feytmans E (1997). Novel knowledge-based mean force potential at atomic level. *J Mol Biol* 267(1): 207-22.

[47] Benkert P, Tosatto SCE, Schomburg D (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins* 71(1): 261-77.

[48] Tosatto S (2005). The victor/FRST function for model quality estimation. *J Comput Biol* 12(10): 1316-27.

[49] Zhou H, Zhou Y (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 11(11): 2714-26.

[50] Samudrala R, Moult J (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* 275(5): 895-916.

[51] Lüthy R, Bowie JU, Eisenberg D. (1992). Assessment of protein models with three-dimensional profiles. *Nature* 356(6364): 83-5.

[52] Jones DT (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* 287(4): 797-815.

[53] Melo F, Sali A (2007). Fold assessment for comparative protein structure modeling. *Protein Sci* 16(11): 2412-26.

[54] Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A (2001). Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci* 10(11): 2354-62.

[55] McGuffin LJ (2007). Benchmarking consensus model quality assessment for protein fold recognition. *BMC Bioinformatics* 8: 345.

[56] Benkert P, Künzli M, Schwede T (2009). QMEAN server for protein model quality estimation. *Nucleic Acids Res* 37(Web Server issue): W510-4.

[57] Moult J (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 15(3): 285-9.

[58] Moult J, Pedersen JT, Judson R, Fidelis K (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins* 23(3): ii-v.

[59] Eidhammer I, Jonassen I, Taylor WR (2005). Protein Bioinformatics: An algorithmic approach to sequence and structure analysis. *Wiley & Sons, Chichester*.

[60] Zemla A, Venclovas C, Moult J, Fidelis K (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins* 37(Suppl 3): 22-9.

[61] Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman JL, Levy Y (2009). Assessment of CASP8 structure predictions for template free targets. *Proteins* 77(Suppl 9): 50-65.

[62] Floudas CA, Fung HK, McAllister SR, Monnigmann M, Rajgaria R (2006). Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science* 61(3): 966-88.

[63] Kryshtafovych A, Fidelis K, Moult J (2007). Progress from CASP6 to CASP7. *Proteins* 69(Suppl 8): 194-207.

[64] Ginalski K (2006). Comparative modeling for protein structure prediction. *Curr Opin Struct Biol* 16(2): 172-7.

[65] Zhang Y (2009). Protein structure prediction: when is it useful?. *Curr Opin Struct Biol* 19(2): 145-55.

[66] Madhusudhan MS, Marti-Renom MA, Eswar N, John B, Pieper U, Karchin R, Shen M, Sali A (2005). Comparative protein structure modelingComparative protein structure modeling. The

Proteomics Protocols Handbook. *JM Walker (ed), Humana Press, Totowa, NJ.*

[67] Ekins S, Mestres J, Testa B (2007). In silico pharmacology for drug discovery: applications to targets and beyond. *Br J Pharmacol* 152(1): 21-37.

[68] Becker OM, Dhanoa DS, Marantz Y, Chen D, Shacham S, Cheruku S, Heifetz A, Mohanty P, Fichman M, Sharadendu A, Nudelman R Kauffman M, Noirman S (2006). An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *J Med Chem* 49(11): 3116-35.

[69] Arakaki AK, Zhang Y, Skolnick J (2004). Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. *Bioinformatics* 20(7): 1087-96.

[70] Ye Y, Li Z, Godzik A (2006). Modeling and analyzing three-dimensional structures of human disease proteins. *Pac Symp Biocomput*: 439-50.

[71] Brylinski M, Skolnick J (2008). Q-Dock: Low-resolution flexible ligand docking with pocket-specific threading restraints. *J Comput Chem* 29(10): 1574-88.

[72] Lengauer T, Rarey M (1996). Computational methods for biomolecular docking. *Curr Opin Struct Biol* 6(3): 402-6.

[73] Raimondo D, Giorgetti A, Bosi S, Tramontano A (2007). Automatic procedure for using models of proteins in molecular replacement. *Proteins* 66(3): 689-96.

[74] Giorgetti A, Raimondo D, Miele AE, Tramontano A (2005). Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics* 21(Suppl 2): ii72-6.

[75] Webb EC (1992). Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. *San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press. ISBN 0-12-227164-5.*

[76] Moult J (2008). Comparative modeling in structural genomics. *Structure* 16(1): 14-6.

[77] Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo JH, Lee J, Baker D, Chivian D, Kim D, Ezkurdia I (2007). Assessment of predictions submitted for the CASP7 domain prediction category. *Proteins* 69(Suppl 8): 137-51.

[78] Zhang Y, Devries ME, Skolnick J (2006). Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput Biol* 2(2): e13.

[79] Malmström L, Riffle M, Strauss CE, Chivian D, Davis TN, Bonneau R, Baker D (2007).

Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. *PLoS Biol* 5(4): e76.

[80] Sippl MJ (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 213(4): 859-83.

[81] Park BH, Levitt M (1996). Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol* 258(2): 367-92.

[82] Park BH, Huang ES, Levitt M (1997). Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J Mol Biol* 266(4): 831-46.

[83] Lazaridis T, Karplus M (1999). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* 288(3): 477-87.

[84] Samudrala R, Xia Y, Huang E, Levitt M (1999). Ab initio protein structure prediction using a combined hierarchical approach. *Proteins* 37(Suppl 3): 194-8.

[85] Gatchell DW, Dennis S, Vajda S (2000). Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* 41(4): 518-34.

[86] Petry D, Honig B (2000). Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Sci* 9(11): 2181-91.

[87] Vendruscolo M, Najmanovich R, Domany E (2000). Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading?. *Proteins* 38(2): 134-48.

[88] Vorobjev YN, Hermans J (2001). ree energies of protein decoys provide insight into determinants of protein stability. *Protein Sci* 10(12): 2498-506.

[89] Dominy BN, Brooks CL (2002). Identifying native-like protein structures using physics-based potentials. *J Comput Chem* 23(1): 147-60.

[90] Felts AK, Gallicchio E, Wallqvist A, Levy RM (2002). Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. *Proteins* 48(2): 404-22.

[91] Wallner B, Elofsson A (2003). Can correct protein models be identified?. *Protein Sci* 12(5): 1073-86.

[92] Zhu J, Zhu Q, Shi Y, Liu H (2003). How well can we predict native contacts in proteins based on decoy structures and their energies?. *Proteins* 52(4): 598-608.

[93] Zhou R, Silverman BD, Royyuru AK, Athma P (2003). Spatial profiling of protein hydrophobicity: native vs. decoy structures. *Proteins* 52(4): 561-72.

[94] McGovern SL, Shoichet BK (2003). Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem* 46(14): 2895-907.

[95] Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M (2004). Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem* 47(1): 45-55.

[96] Murray CW, Baxter CA, Frenkel AD (1999). The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J Comput Aided Mol Des* 13(6): 547-62.

[97] Vakser IA (1996). Low-resolution docking: prediction of complexes for underdetermined structures. *Biopolymers* 39(3): 455-64.

[98] Wojciechowski M, Skolnick J (2002). Docking of small ligands to low-resolution and theoretically predicted receptor structures. *J Comput Chem* 23(1): 189-97.

[99] Bindewald E, Skolnick J (2005). A scoring function for docking ligands to low-resolution protein structures. *J Comput Chem* 26(4): 374-83.

[100] Schafferhans A, Klebe G (2001). Docking ligands onto binding site representations derived from proteins built by homology modelling. *J Mol Biol* 307(1): 407-27.

[101] Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature* 450(7167): 259-64.

[102] Taylor WR (2006). Decoy models for protein structure comparison score normalization. *J Mol Biol* 357(2): 676-99.

[103] Fetrow JS, Skolnick J (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol* 281(5): 949-68.

[104] Wei L, Huang ES, Altman RB (1999). Are predicted structures good enough to preserve functional sites?. *Structure* 7(6): 643-50.

[105] Nayal M, Di Cera E (1994). Predicting Ca2+ binding sites in proteins. *Proc Natl Acad Sci U S A* 91(2): 817-21.

[106] Yamashita MM, Wesson L, Eisenman G, Eisenberg D. (1990). Where metal ions bind in proteins. *Proc Natl Acad Sci U S A* 87(15): 5648-52.

[107] Samudrala R, Levitt M (2000). Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Sci* 9(7): 1399-401.

[108] Zhang Y, Kihara D, Skolnick J (2002). Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* 48(2): 192-201.

[109] Bork P, Ouzounis C, Sander C (1994). From genome sequences to protein function. *Curr. Opinion in Struct. Biol. 4,* 4(3): 393–493.

[110]Koonin EV, Tatusov RL, Rudd KE (1996). Protein sequence comparison at genome scale. *Methods Enzymol* 266: 295-322.

[111]Galperin MY, Koonin EV (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol* 1(1): 55-67.

[112]Todd AE, Orengo CA, Thornton JM (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 307(4): 1113-43.

[113]Barrett AJ (1997). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur J Biochem* 250(1): 1-6.

[114]Devos D, Valencia A (2000). Practical limits of function prediction. *Proteins* 41(1): 98-107.

[115]Rost B (2002). Enzyme function less conserved than anticipated. *J Mol Biol* 318(2): 595-608.

[116]Tian W, Skolnick J (2003). How well is enzyme function conserved as function of pairwise sequence identity?. *J Mol Biol* 333(4): 863-82.

[117]Wilson CA, Kreychman J, Gerstein M (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* 297(1): 233-49.

[118]Zhang Y, Kolinski A, Skolnick J (2003). TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys J* 85(2): 1145-64.

[119]Zhang Y, Skolnick J (2004). Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci U S A* 101(20): 7594-9.

[120] Chelliah V, Chen L, Blundell TL, Lovell SC (2004). Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J Mol Biol* 342(5): 1487-504.

[121] Taylor WR (2002). A 'periodic table' for protein structure. *Nature* 416(6881): 657-60.

[122] Chelliah V, Taylor WR (2008). Functional site prediction selects correct protein models. *BMC Bioinformatics* 9(Suppl 1): S13.

[123] Wang G, Dunbrack RL Jr (2003). PISCES: a protein sequence culling server. *Bioinformatics* 19(12): 1589-91.

[124] Zemla A (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31(13): 3370-4.

[125] Schrödinger LLC. The PyMOL Molecular Graphics System, Version 1.2r3pre.

[126] Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/.

[127] Schreyer A, Blundell T (2009). CREDO: a protein-ligand interaction database for drug discovery. *Chem Biol Drug Des* 73(2): 157-67.

[128] Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002). Analysis of catalytic residues in enzyme active sites. *J Mol Biol* 324(1): 105-21.

[129] Porter CT, Bartlett GJ, Thornton JM (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32(Database issue): D129-33.

[130] Kabsch W, Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12): 2577-637.

[131] Laskowski RA (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 13(5): 323-30, 307-8.

[132] Mihel J, Sikić M, Tomić S, Jeren B, Vlahovicek K (2008). PSAIA - protein structure and interaction analyzer. *BMC Struct Biol* 8: 21.

[133] Lee B, Richards FM (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 55(3): 379-400.

[134] Shrake A, Rupley JA. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol* 79(2): 351-71.

[135] Phillips SEV, Moras D (1995). Protein-nucleic acid interactions. *Curr Opin Struct*

Biol 5(1): 1-3.

[136] Weber G (1992). Protein Interactions. *Chapman and Hall, New York*.

[137] Zvelebil MJ, Thornton JM (1993). Peptide-protein interactions: An overview. *Q Rev Biophys* 26(3): 333-63.

[138] Janin J, Chothia C (1990). The structure of protein-protein recognition sites. *J Biol Chem* 265(27): 16027-30.

[139] Jones S, Thornton JM (1995). Protein-protein interactions: A review of protein dimer structures. *Prog Biophys Mol Biol* 63(1): 31-65.

[140] Kuntz ID, Meng EC, Shoichet BK (1994). Structure-based molecular design. *Acc Chem Res* 27(5): 117–123.

[141] Colman PM (1994). Structure-based drug design. *Curr Opin Struct Biol* 4(6): 868-74.

[142] Pintar A, Carugo O, Pongor S (2002). CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics* 18(7): 980-4.

[143] Richards FM (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol* 82(1): 1-14.

[144] Pintar A, Carugo O, Pongor S (2003). DPX: for the analysis of the protein core. *Bioinformatics* 19(2): 313-4.

[145] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB (2003). Protein disorder prediction: implications for structural proteomics. *Structure* 11(11): 1453-9.

[146] Goldenberg O, Erez E, Nimrod G, Ben-Tal N (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res* 37(Database issue): D323-7.

[147] Armon A, Graur D, Ben-Tal N (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* 307(1): 447-63.

[148] Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19(1): 163-4.

[149] Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005).

ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* 33(Web Server issue): W299-302.

[150] Karplus PA, Schulz GE (1985). Prediction of chain flexibility in proteins - a tool for the selection of peptide antigens. *Naturwissenschaften* 72: 212–3.

[151] Vihinen M, Torkkila E, Riikonen P (1994). Accuracy of protein flexibility predictions. *Proteins* 19(2): 141-9.

[152] Vihinen M (1987). Relationship of protein flexibility to thermostability. *Protein Eng* 1(6): 477-80.

[153] Parthasarathy S, Murthy MR (2000). Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng* 13(1): 9-13.

[154] Carugo O, Argos P (1998). Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins* 31(2): 201-13.

[155] Yuan Z, Zhao J, Wang ZX (2003). Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng* 16(2): 109-14.

[156] Mohan S, Sinha N, Smith-Gill J (2003). Modeling the binding sites of anti-hen egg white lysozyme antibodies HyHEL-8 and HyHEL-26: an insight into the molecular basis of antibody cross-reactivity and specificity. *Biophys J* 85(5): 3221-36.

[157] Carugo O, Argos P (1997). Correlation between side chain mobility and conformation in protein structures. *Protein Eng* 10(7): 777-87.

[158] Eyal E, Najmanovich R, Edelman M, Sobolev V (2003). Protein side-chain rearrangement in regions of point mutations. *Proteins* 50(1): 272-82.

[159] Altman R, Hughes C, Zhao D, Jardetsky O (1994). Compositional characteristics of relatively disordered regions in proteins. *Prot Pept Letters* 1: 120-127.

[160] Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK (2004). Protein flexibility and intrinsic disorder.. *Protein Sci* 13(1): 71-80.

[161] Navizet I, Lavery R, Jernigan RL (2004). Myosin flexibility: Structure domains and collective vibrations. *Proteins* 54(3): 384-93.

[162] Tronrud DE (1996). Knowledge-based B-factor restraints for the refinement of proteins. *J Appl Crystallogr* 29: 100-4.

[163] Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G (2003). Improved amino acid flexibility parameters. *Protein Sci* 12(5): 1060-72.

[164] Parthasarathy S, Murthy MR (1997). Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci* 6(12): 2561-7.

[165] Diella F, Gould CM, Chica C, Via A, Gibson TJ (2008). Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res* 36(Database issue): D240-4.

[166] Zanzoni A, Ausiello G, Via A, Gherardini PF, Helmer-Citterich M (2007). Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res* 35(Database issue): D229-31.

[167] Zanzoni A, Carbajo D, Diella F, Gherardini PF, Tramontano A, Helmer-Citterich M, Via A (2011). Phospho3D 2.0: An enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic Acids Res* 39(Database issue): D268-71.

[168] Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 10: 136.

[169] The UniProt Consortium (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 39(Database issue): D214-9.

[170] Miller S, Janin J, Lesk AM, Chothia C (1987). Interior and surface of monomeric proteins. *J Mol Biol* 196(3): 641-56.

[171] Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982). A geometric approach to macromolecule-ligand interactions. *J Mol Biol* 161(2): 269-88.

[172] Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N (2005). The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins* 58(3): 610-7.

[173] Durek P, Schudoma C, Weckwerth W, Selbig J, Walther D (2009). Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics* 10: 117.

[174] Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 8(11): R250.

[175] Gnad F, Gunawardena J, Mann M (2011). PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* 39(Database issue): D253-60.