Chapter 9

On the Assessment of Structural Protein Models with ROSETTA-Design and HMMer: Value, Potential and Limitations

León P. Martínez-Castilla and Rogelio Rodríguez-Sotres

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/47842

1. Introduction

The prediction of the three-dimensional structure of a protein, starting with the amino acid sequence, is still an unsolved issue. However a number of important advancements have been made and some methods offer solutions to this problem, specially when the target sequence has homologues whose structure has been determined. In any case, it is important to evaluate the quality of the prediction, as none of the methods offers assurance of success. The ROSETTA-design-HHMer (Rd.HMM) protocol stands out among the current quality assessment methods, because it offers evidence of the biological appropriateness of the prediction. In addition, Rd.HMM can be used to guide the modeling process towards the improvement of the model's quality. This chapter deals with the principles behind this protocol and gives practical advice on how to use the Rd.HMM to evaluate the quality of a three-dimensional modeled structure of a protein, and how to use the information to improve the model. The limitations of the protocol are also discussed.

2. The folding problem is a NP-hard problem involving a degenerate informational code

As implied by the well-known Levinthal paradox (Levinthal, 1968), a full exploration of the entire conformational space theoretically available to a protein is out of the reach of current computational techniques. Equally unaccessible to nature is the sequence space available to polypeptide chains (Kono & Saven, 2001). Currently, the amount of available protein structures (the PDB) represents a fraction of the known protein amino acid sequences, and if the available sample is grouped in terms of different folds, the diversity in the PDB is even smaller. In addition, protein structure and function can tolerate a significant number of

mutations. Both facts suggest an important degree of degeneracy between the information in polypeptide sequences and the associated code leading to their native structure (Bowie et al., 1990). In other words, the so-called folding code is degenerate.

However, even if the number of protein structural folds is smaller that the sequence space, the folding problem is still unsolved, because exploring the total number of conformations available to a protein or its energy landscape are NP-hard problems (Hart & Istrail 1997), and because the available methods to calculate the energy of a protein conformation imply a large systematic error (Faver et al., 2011).

The above facts set forth the intractability of solving the problem through an exhaustive search. Nevertheless, proteins in nature do reach a native structure in short times, and finding a native-like solution to the three-dimensional structure of a protein may not require a full examination of the conformational space, or its corresponding energy landscape. In fact, recent years have seen important progress in the search for solutions to the protein folding problem (Dill et al., 2008).

3. The problem of quality scoring for 3D models

In theory, the native three-dimensional structure of a protein must lie at an energy minimum, underneath all accessible intermediates with near-native fold. However, an accurate calculation of the energy for a protein conformation requires quantum chemical calculations. Properties such as electron-electron correlation, charge transfer, polarization, and bond break/formation, including proton exchange, involve quantum mechanical effects and cannot be correctly described using the equations of classical physics. The relevance of quantum mechanics for accurate energy calculations of protein-ligand complexes and protein conformations have been recently demonstrated (Raha and Merz, 2005). Numerical approximations to the electronic state of a multielectronic system have been developed for a variety of system up to date. But only a few simplified solutions, implying low-precision, can tackle an electronic macromolecular system, and even these demand a large amount of computational resources (He & Merz, 2010). The common simplifications, based on molecular mechanics, do carry a systematic error that precludes the accurate finding of the true native energy minimum (Faver et al., 2011).

Many methods have been proposed to model the three-dimensional structure of proteins starting from their amino acid sequence. Based on their use of experimental structural information, these methods can be classified into comparative modeling or *ab initio* methods.

Because rating the success of any method requires an impartial judge to be trustworthy, the scientific community implemented the contests for CRITICAL ASSESMENT OF THE STRUCTURE OF PROTEINS (CASP) (Kryshtafovych et al., 2009). In such contests, the judges are computer algorithms, which compare a 3D-structure solved by an experimental method (but yet unpublished) to a 3D-model predicted by a CASP contestant. The comparative modeling strategies have had a remarkable degree of success in the prediction of 3D-structures of soluble proteins, with the amino acid sequence as starting information.

Comparative modeling exploits the wealth of experimental structural information nowadays available for proteins (Rose et al., 2011), and relies on powerful sequence alignment algorithms (Wallace et al., 2005). In CASP contests, comparative modeling servers, such as I-TASSER (Roy et al., 2010), ROBETTA (Kim et al., 2004) and SAM-T08 (Karplus, 2009), have achieved a high success rate in their predictions for protein 3D-structures of low to intermediate difficulty (as defined by the CASP staff). Yet, one mayor limitation in these methods lies in the strategies used to match each amino acid in a target sequence to its corresponding best hosting spot in the 3D-structure of the template and, again, this is a NP-complete problem (Lathrop, 1994).

In *ab initio* methods, the laws of physics and chemistry and/or artificial intelligence are used to generate a prediction for a native-like folding solution of a protein with known amino acid sequence (Dill et al., 2008). While *ab initio* methods have been less successful than comparative modeling, these are the only choice if no suitable homologous 3D-template is available, for a given amino acid sequence (Kryshtafovych et al., 2009).

The above considerations are all fine when the question is to grade the methods and chose the one with highest success rate, but to date, no single method gives the correct answer every time. Yet, the final aim of such methods is to produce good native-like protein 3D-predictions, when experimental X-ray or NMR data are not available. How then is it possible to set apart models with wrong fold assignment, from those with a correct fold assignment, but with a mistraced sequence to 3D-fold alignment (Luthy et al., 1992)? Is it possible to identify cases where the fold assignment and the alignment are adequate, but the solution to the atom repacking of replaced amino acids is deficient? These questions lie behind the quality assessment of a protein 3D-structure prediction.

The quality assessment is of particular relevance in cases where a suitable 3D-template cannot be found, because the predicted 3D-model cannot be compared back the starting template. Again, this problem can be tackled with a number of strategies, and most of them have been implemented as computer software programs, and their validity tested at the CASP contests (Shi et al., 2009).

Quality assessment methods for the predicted 3D-structures of proteins can be classified according to their underlying principles:

i. Physics-based methods use the regularities in chemical structures and the laws of physics and chemical bonding to find how much a 3D-structure deviates from the known canonical values. These methods may come in the form of force-fields and they report energies (Hu & Jiang, 2010), or may seek for abnormalities in geometrical and chemical features such as bonding lengths, bonding angles, dihedral torsion values, charge-charge distances and so on (Rodriguez et al. 1998).
ii. Statistics-based methods use the known 3D-structures to generate a set of probability distributions for a number of features of the experimentally solved structures. These distributions can be used as reference to judge the quality of a prediction. When these probability distributions are transformed into energies, using the Boltzmann law, the

result is designated as a statistical potential. Although statistical potentials started as empirical constructs, their theoretical basis have been substantiated recently (Hamelryck et al., 2010). These constructs turned out be very useful since any experimental quantitative variable can be treated as an energy and used to generate a potential landscape for 3D-structures. Amongst these latter methods, ANOLEA (Melo & Feytmans, 1998) has a simple conception, and it can be calculated quickly and with a modest computer system, even for very large 3D-structures. Despite its simplicity, ANOLEA stands as one of the most reliable quality assessment indices (Chodanowski et al., 2008).

iii. Artificial intelligence programs such as neural networks, or support vector machines have shown limited success in predicting the 3D-structure of proteins, but their success in quality assessment has been acceptable. A number of these programs has appeared through the years and, again, these methods depend on experimental data to train or setup the program's intelligence (Wallner & Elofsson, 2003). Unfortunately, what features has the computer learned to judge is not always clear, and in some specific cases, the results may be unexpected.

iv. Finally, hybrid methods combine different strategies to test a 3D-structure quality. Amongst these methods, web metaservers, such as metaMQAP (Pawlowski etal., 2008), deserve a note, because they meld the scores from a number of servers into a weighted quality index of a 3D-structure.

While most methods mentioned above may be of value to assess the quality of a predicted protein 3D-structure, it is possible for a model to have acceptable geometrical features, resemble the fold of a structure in the PDB, and still represent a non-native 3D-conformation of the protein under consideration. We have designated this limitation as the appropriateness problem of a 3D-structure prediction. After a careful analysis of several related methods, in our opinion, only the recently published protocol ROSETTA-design-HMMer (Rd.HMM) (Martínez-Castilla & Rodríguez-Sotres 2010) offers robust and explicit evidence of the biological appropriateness of a protein 3D-structure.

4. The reverse folding problem

Due to the degeneracy of the amino acid sequence to three-dimensional fold translation code (Bowie et al., 1990), discussed above, proteins can tolerate amino acid changes in their sequence, as long as these changes do not fall in positions crucial to their folding stability, folding kinetics, macromolecular meaningful interactions, conformational transitions, ligand binding, or catalytic function. Therefore, two proteins sharing more than 40% sequence identity are likely to participate in the same or very similar cellular functions. Based on these considerations, sequence databases may be automatically annotated based on sequence homology between the new unannotated entries and already annotated ones.

As an additional consequence of the folding code degeneracy, the prediction of a 3D-fold starting with the amino acid sequence, *i.e.* the folding problem, is a far more complex

problem than it is the reverse folding problem, which attempts to predict an amino acid sequence compatible with the atomic 3D-coordinates of a protein backbone. One of the first approaches to this problem was published by Eisemberg and co-workers (Luthy et al., 1992; Wilmanns & Eisenberg, 1995). According to their data, given a set of the atomic 3D-coordinates from the native 3D-structure of a protein, it is possible to reconstruct the amino acid sequence of the corresponding natural protein, with a good level of confidence.

A second attempt was published by the group of David Baker (Cheng et al., 2005), who expanded the search beyond the natural amino acid sequence of the protein, to explore part of the sequence space compatible with a given 3D-fold. These authors used the 3D-atomic coordinates from a protein backbone to complement the set of amino acid sequences from natural homologues, with a set of predicted artificial amino acid sequences. In the alignment from this set, they could distinguish the conservation due to structural constraints from the functional conservation. Their data indicated a clear tendency of functional sites to have sub-optimal free energies of stability and their computed sequence profiles diverged from the natural sequence profile. This method was offered as a web service to predict functional sites (Protinfo MFS, http://protinfo.compbio.washington.edu/mfs/, accesed on may 15, 2012).

In a later work, Chivian and Baker (Chivian & Baker, 2006) used a sophistication of the earlier approach to refine a sequence-to-structure alignment, as part of an homology modeling protocol. Their data showed an increase in the alignment's quality of a target amino acid sequence to a 3D-template. These authors integrated this alignment method in the ROBETTA 3D-structure prediction server (Kim et al., 2004). As mentioned in the preceding section, ROBETTA has been repeatedly among the top servers in recent CASP contests and, very likely, this alignment method is part of its success.

In the approaches discussed in this section, the authors applied strategies to account for the conformational flexibility of the backbone in their search, widening the range of amino acid choices for these segments. Therefore, the higher the backbone flexibility, the lower the conservation and the higher the likelihood of such site to be declared as functional. In addition, during the estimation a region's flexibility, part of the natural amino acid information must be retained, because the instability of any segment is intrinsically linked to the properties of the local side chains and their neighbors.

The alternative to this search is to accept the 3D-coordinates for the X-ray solved structure as valid equilibrium conformations, and ignore those segments where the excessive mobility prevented the assignment of atom positions. In NMR solved structures, there is usually more information on accessible conformations, and the approach may take this into account, or use the more populated conformation. In this last case, the conformational flexibility is lost, but the computed set of sequences will make a better sampling of the sequence space available to this particular equilibrium conformation.

From this considerations, any attempt to explore the sequence space available to a given fold clearly must accept some informational loss, but at this point, the sequence space compatible with a completely fixed backbone was in need of a deeper exploration.

In the Rd.HMM protocol (Martínez-Castilla & Rodríguez-Sotres 2010), ROSETTA-design (Rd) is used to redesign the 3D-structure of a protein by reassigning amino acids to every position in the structure, and with no restriction in the choice of amino acids or rotamers. To completely suppress the information present in the starting amino acid sequence, a preliminary redesign of the protein is made by imposing to the 3D-backbone a fixed new random sequence. To reduce any bias possibly introduced by this random sequence, this step is performed several times. When scored with the ROSETTA force-field for stability, the 3D-structures with randomized sequence have very high energies, because the artificial side chains will frequently fail to fit into the cavities left by the natural side chains, and neighboring contacts are likely to be unfavorable. In other words, these randomized sequence 3D-models are *in silico* constructs, meaningless in terms of chemistry or biology.

In the second step, Rd is used to redesign each 3D-structure with randomized sequence produced before, but this time with complete freedom of amino acid choice, and the reconstruction is done many times. Rd can be trusted to find amino acids combinations with high stability (Kuhlman et al., 2003; Jiang et al., 2008; Slovic et al., 2004; Butterfoss et al., 2006; see also next section) and each new redesign will harbor a new theoretically low-energy sequence of amino acids for the 3D-backbone under consideration, but most likely, a non-natural one, because the selection pressure in natural proteins is not limited to stability constraints (Cheng et al., 2005).

In the end, a set of amino acid sequences can be recovered from the corresponding set of 3D-redesigns, as large as requested, and representing a sample of theoretically possible, but naturally inexistent amino acid combinations, optimized only for 3D-fold stability. The theoretical stability of the redesigns are expected to exceed natural protein stability (Cheng et al., 2005; Butterfoss et al., 2006), but a folding pathway to the 3D-fold may not exist for such sequences, because ROSETTA-design has not been imprinted with any information related to the folding process. That is to say, no all redesigns are expected to fold correctly in experimental tests.

5. The merits of ROSETTA-design

ROSETTA-design (Rd) is a program developed by the group of David Baker (Kuhlman et al., 2003) with a remarkable success in the design of suitable amino acid sequences for a given-fold. The ROSETTA suite includes modules for protein structure refinement, a*b initio* protein folding predictions, antibody design, protein-ligand docking, protein-protein docking, and others. However the merits and limitations of those other protocols will not be discussed here.

Rd was created with one application in mind, namely "to find amino acid sequences able to fold into a given three-dimensional structure". To this aim, Baker's group developed three basic components: a modified force-field with a large penalty for atomic overlap, a rotamer database taken from the PDB and refined with quantum chemical calculations, and a Monte-Carlo search algorithm to replace the amino acid side-chains of the starting structure (Kuhlman et al., 2003).

The approach followed by Rd has proven very robust because it made possible to design the first artificial protein folding into a completely novel topology (Kuhlman et al., 2003). Rd has been also used with success to place a novel enzyme active site, of human design, into an unrelated protein (Jiang et al., 2008), and to convert a membrane protein into a soluble protein (Slovic et al., 2004), among other notable protein engineering applications (Butterfoss et al., 2006).

Monte-Carlo methods can be implemented in algorithms to various aims. Some are designed to provide an extensive sampling of a given landscape, but in other cases the algorithm is set to find a optimum (usually a minimum) in such landscape. The very well-known Metropolis algorithm (Metropolis et al., 1953) can be used for both purposes, but it has been theoretically proven to converge to the true optimum, if no time limit is set (Mengersen & Tweedie, 1966). In practice, Monte-Carlo methods may take too many steps and the search has to be stopped when the sampling is considered extensive enough, usually, well before the true optimum is determined (Cowles & Carlin, 1996).

Once again, due to the degeneracy in the folding code (see section 1), low-energy solutions for amino acid side chain replacements on a 3D-backbone have many local minima, and some may be within the reach of a short to moderate Monte-Carlo random-walk. Rd narrows down the list of amino acid rotamers to be tried at each α carbon, uses a computer-efficient code for energy calculations, an improved force-field, and has a curated database of rotamers, with improved geometries obtained through quantum mechanical calculations. In addition, Rd starts with a geometrical analysis of the structure and removes from the search amino acid sites where the local environment makes the choices' list too narrow or too undefined. The assignment at those sites becomes then trivial.

Finally, Rd can be fed with a list of amino acid choices for each residue in the 3D-backbone, ranging from not allowing changes, to the full set of 20 amino acids and all of their rotamers. Rd is, therefore, one of the most flexible programs for protein design (Butterfoss et al., 2006).

6. Hidden Markov models to deal with the reverse folding problem

A Markov model (MM) is a model of a stochastic process with the Markov property. The model has the Markov property if, along the random succession of states, the future state is determined by the present state only, with no influence of the previous states (Eddy, 2004). The change from one state to another is called a transition, and each state has an associated transition probability. The states are finite or countable, but the succession itself may be infinite.

A MM can be used to describe a number of natural phenomena. For example, in a chemical kinetic mechanism, the states are chemical intermediates and transition probabilities are rate equations (Shapiro & Zeilberger, 1982). When these states constitute symbol emitters and each state has a defined emission probability for each possible symbol, and a concatenation of states will broadcast a symbols' sequence, for instance, an amino acid sequence (Eddy et al., 1995).

A very simple sequence-generating MM may consist of two states (Fig. 1): Let the state S_1 be an emitter of any of the 20 amino acids abbreviations. Let the amino acid compositions of an infinite sequence of symbols produced by S_1 equal the composition of natural proteins. With a probability of 0.1, state S_1 may suffer a transition to a second emitter S_2. In turn, S_2 is able to emit a stop, or to transit back to S_1, with 0.9 probability. This two states will go forth and back to give an infinite number of sequences of short length, because, given the probabilities, sequences longer than a few tenths of amino acids will be very infrequent.

Since a MM is a stochastic device, it is unsuited to represent only one particular sequence, but instead, it can be a powerful tool to represent a subset of the sequence space, notably, a sequence alignment. Such MM represents the observed aligned sequences, usually a subset of all the possible sequences in the alignment, but the states of the model (each one encoding the probabilities of one or more alignment positions) cannot be observed. When such is the case, the MM is then said to be hidden (HMM). However, the Viterbi algorithm, the forward algorithm, and the Baum–Welch algorithm make it possible to compute the most likely parameters of the model's states, out of the observations available (Eddy, 2004; Eddy et al., 1995).

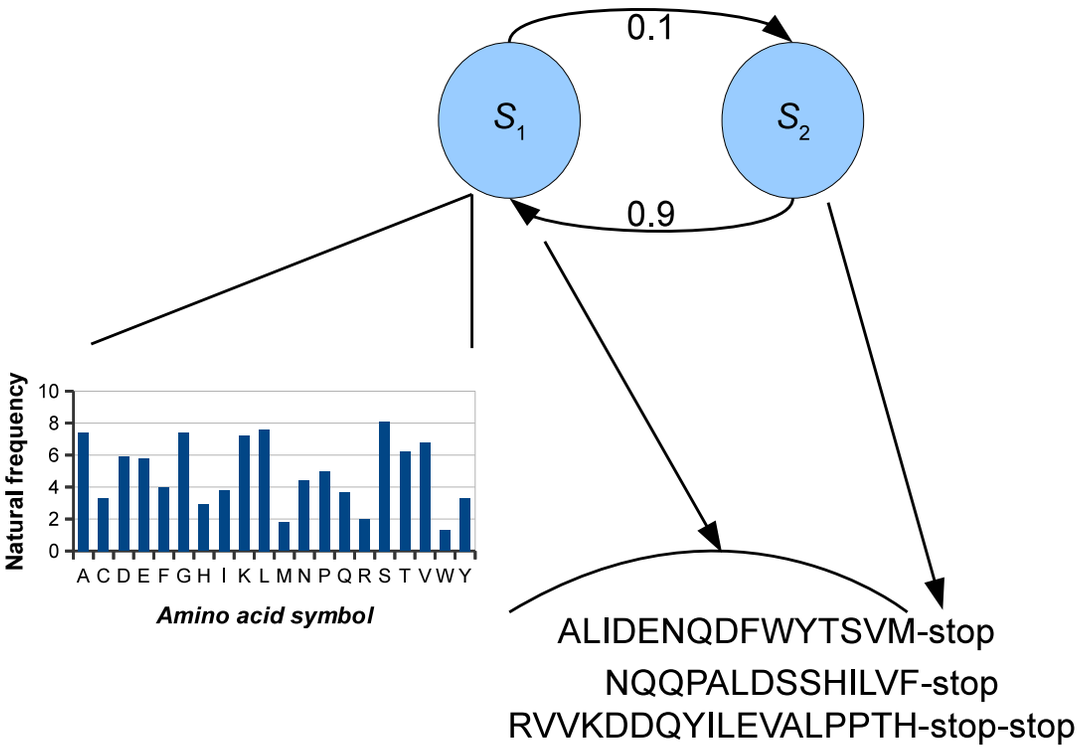


Figure 1. A simple Markov model to emit random amino acid sequences of variable length with an amino acid composition similar to that in natural proteins.

Because the HMM represents more sequences than those observed, it can be used to produce new sequences, but most importantly, for any available sequence in a database, its emission probability by the HMM can be calculated and compared to corresponding

emission probability by a very general model, such as the one in figure 1. The ratio of these two probabilities can be used as an index or score, the higher the score, the higher the likelihood of the sequence being a member of the alignment. From this information, the expectancy of such an index value being due to chance can be estimated. Expectancies of one or above may indicate a meaningless score.

HMMer is a suite of programs developed by Sean Eddy (Eddy, 2004, Eddy et al., 1995) to create and use HMMs of amino acid and nucleic acid sequence alignments. HMMer has executables to estimate the parameters of a HMM from a sequence alignment and calibrate the model to allow a good estimation of scores and expectancies. Other executables will test a sequence database to extract those sequences with high score and low expectancy, aligning the new sequences to the model. Additional executables can create the starting HMM, use it to emit sequences with high probability of being members of the model, or update the model parameters using the newly discovered additional sequences.

One critical step in the HMM preparation is the starting alignment fed to HMMer (Eddy, 2004), because as mentioned in section 1, the optimal alignment of sequence sets is not a trivial problem (Lathrop, 1994). When a HMM gives poor results, it is frequently as a consequence of a defective alignment.

Another limitation of a HMM lies on its very definition, because a MM must be memoryless (Markov property). In the 3D-structure of proteins those amino acids brought into proximity during folding, must be of compatible nature from the sterical and chemical points of view. This property is stored in the sequence as sites with correlated variability, also known as mutual information. Its relevance has been recognized and exploited (Socolich et al., 2005), but HMM are unable to encode such information.

After the above discussion of HMMer features, we can consider its value in dealing with the large a set of amino acid sequences redesigned by the Rd protocol described in the preceding section:

1. Rd.HMM produces many sequences that can be trivially aligned, because every amino acid has biunivocal correspondence with a 3D-backbone site.
2. Using a HMM to represent the redesigned sequences will result in the statistical extension of the sample to sequences with a similar frequency profile. This extension is however inaccurate, because not all off the sequences possibly emitted by the HMM will actually be low-energy solutions to the 3D-backbone redesign (Hamelryck et al., 2010).
3. The HMM can be used to search those natural sequences having amino acid combinations suitable to the 3D-structure under analysis. The value of HMM in the analysis of relationships between biological sequences has been extensively documented (Eddy, 2004).
4. Due to (1), the search made in a database of natural sequences by means of the HMM will align each selected sequence in a structurally aware manner (Martínez-Castilla & Rodríguez-Sotres 2010). But given (2), such structurally aware alignment is somehow inaccurate, the lower the HMMer score, the less reliable this alignment becomes.

7. The unexpected sensitivity of Rd-HMMer

In theory, when a Rd.HMM is used to scan a general sequence database, such as the NCBI-nr (Jiang et al., 2008), a sequence is selected if it is considerably less likely to be generated at random, than to be emitted by the Rd.HMM. But the Rd-step leaves only information related to the 3D-fold, which is then fed into the HMM, thus any selected sequence should be able to fold into a 3D-structure very similar to the starting one. Sequences selected by HMMer should then belong to the same folding family.

One of the unexpected results of Rd.HMM is the sensitivity of this protocol, for instance, it is able to separate those sequences of the TIM-barrel fold that belong to the triose phosphate isomerase from those that belong to other TIM-barrels, such as the phosphoribosylpyrophosphate isomerase (PRAI) (Martínez-Castilla & Rodríguez-Sotres 2010).

Apparently, the Rd-step can imprint its artificial sequences with some details related to loop and turn shapes, as well as contact between secondary structure elements within the tertiary structure adopted by the original polypeptide chain. Then, only when two proteins with completely different activity retain an almost identical structure, a single Rd.HMM can score their corresponding sequences with a significant score. Such is the case of the novel engineered retroaldolases (RA-61, RA-22) and the corresponding templates used to host the newly designed amino acid catalyst, a β-1,4-*endo*-xylanse from *Nonomuraea flexuosa* and one indole-3-glycerol phosphate syntase from *Sulfolobus solfataricus* (Martínez-Castilla & Rodríguez-Sotres 2010; Jiang et al., 2008). This was also the case with the imidazoleglycerolphosphate synthase From *Thermotoga maritima* and the engineered imidazoleglycerol_evolvedcerolphosphate synthase (Martínez-Castilla & Rodríguez-Sotres 2010; Röthlisberger et al., 2008).

The remarkable sensitivity of the Rd.HMM protocol is reflected also in the change of the score reported for the 3D-structure of one protein resolved by NMR, as compared to its X-ray 3D-structure. An Rd.HMM produced with a X-ray 3D-structure will score its corresponding natural sequence with a value close to 0.6 times the length of its amino acid sequence. Instead a Rd.HMM from an NMR derived structure will report half of that score for its corresponding natural sequence (Martínez-Castilla & Rodríguez-Sotres 2010).

As an additional test of the Rd.HMM sensitivity, we compared the Rd.HMMs corresponding to subunit A from two prokaryotic glycyl-tRNA-synthases, one from *Thermus thermophilus* and another from *Thermotoga maritima.* These two X-ray resolved structures have a very similar core (Fig 2A), but the sequence similarity is below 15%. Despite the structural similarity, both proteins have extensive regions where the structure differs completely. Accordingly, the 1ATI:A Rd.HMM scored the *T. thermophilus* sequence (its corresponding natural one) with a value of 161.8 (score over sequence length 0.37) and a highly significant *E*-value (3.8×10^{-49}), but the *T. maritima* sequence received a negative score of -271.1 (score over sequence length -0.94) lacking statistical significance (*E*-value 2). In contrast, the 1J5W:A Rd.HMM scored the *T. thermophilus* sequence with a value of -200.0 (-

0.88) lacking significance (E-value 0.065), and the *T maritima* sequence (its corresponding natural one) received a positive score of 30.3 (0.11) and high statistical significance (*E*-value 8.4×10^{-17}). In these cases, the score was obtained lowering the software threshold, because in a standard search of the NCBI-nr the 1ATI:A Rd.HMM only identified the *T. thermophilus* glycyl-tRNA amino acid sequence and its homologues.

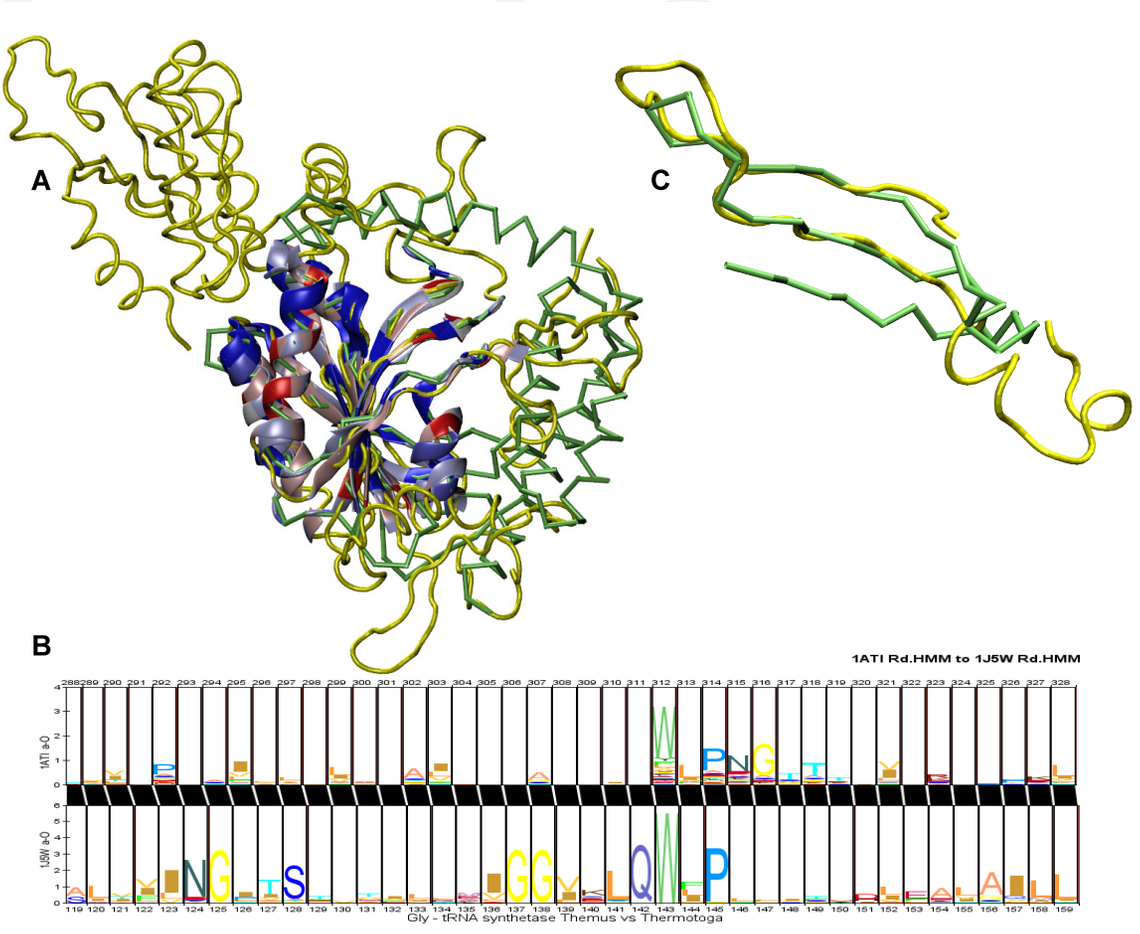


Figure 2. (A) Comparison of glycyl-tRNA synthetases from *Thermus thermophilus* (PDB 1ATI:A, yellow tube) and from *Thermotoga maritima* (PDB 1J5W:A, green trace). The core α/β region was superimposed using TOPOFIT (Ilyin et al., 2004) (shown as cartoons) and colored according to its sequence similarity from blue (identical) to white (dissimilar). The figure was prepared using VMD (Humphrey et al., 1996). (B) HMM logo of the profile to profile alignment (Schuster-Böckler & Bateman, 2005) of Rd.HMMs from glycyl-tRNA synthetases in (A). (C) The segments in (A) corresponding to the nodes in the alignment in (B)

In the previous example, the dissimilar regions have enough information to allow the discrimination between the structures. In addition, since the scores for the non-related sequence on each case were negative, the alignment produced by the Rd.HMM of both sequences is unreliable. Figure 2B shows the profile to profile comparison of HMM logos (Schuster-Böckler & Bateman, 2005) for the Rd.HMM derived from both glycyl-tRNA synthetases, which paired a significant subset of both Rd.HMMs. The corresponding segments were indeed structurally related (Fig. 2C).

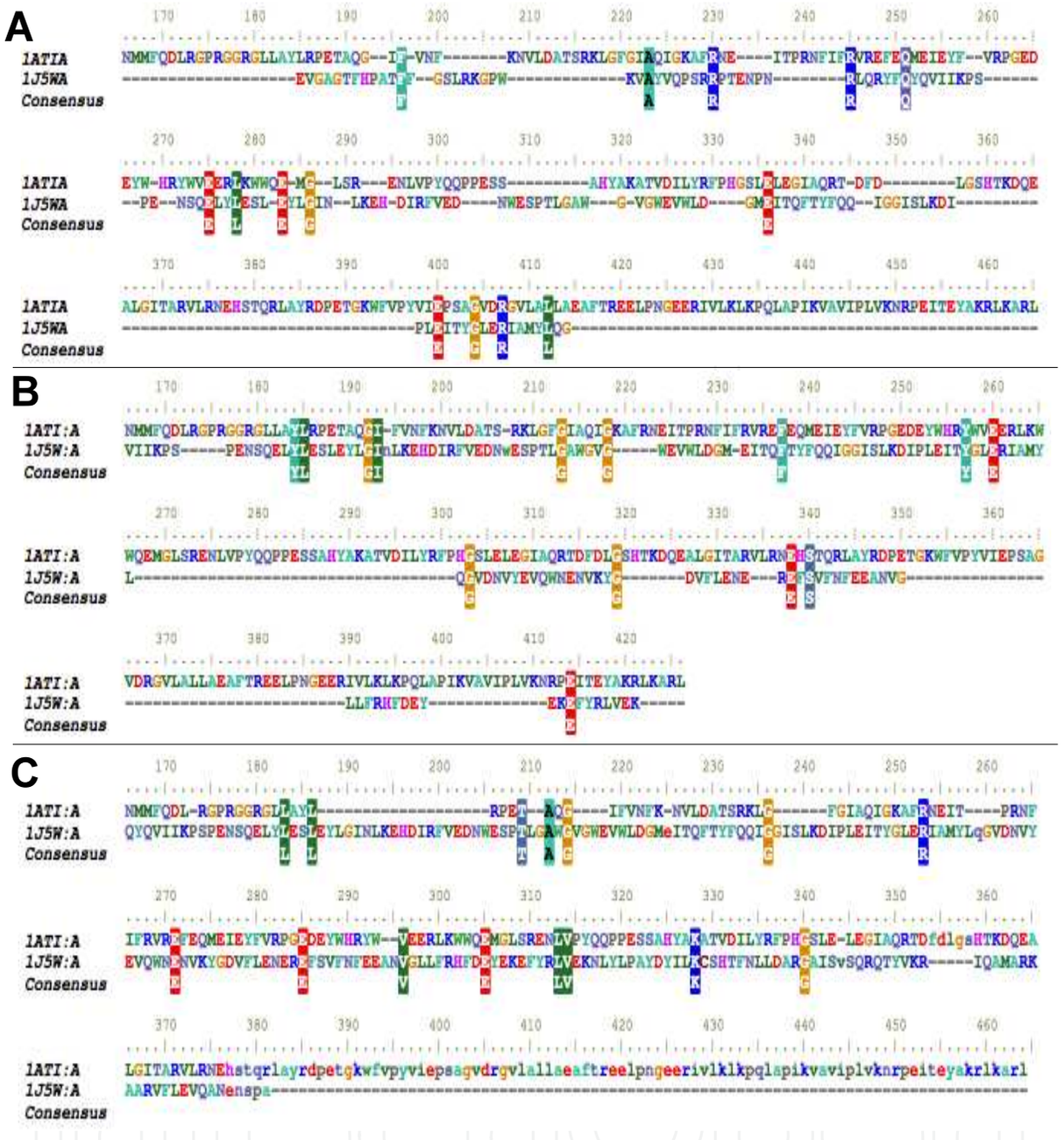


Figure 3. (A) TOPOFIT (Ilyin et al., 2004) sequence alignment based on the structural alignment in figure 2. Amino acids are aligned if their backbones are less than 3 Å apart. (B) Alignment guided by the Rd.HMM derived from 1ATI:A. (C) Alignment guided by the Rd.HMM derived from 1J5W:A. For clarity, only the section of the alignment including aminoacids 166 to 424 of 1ATI:A is shown.

Figure 3 (A to C) shows the lack of coincidence between the TOPOFIT structural alignment (Ilyin et al., 2004), and the two Rd.HMM based alignments for the core regions. A careful analysis of the alignments in figure 3 suggests a possible explanation for the notable specificity of 1ATI:A and 1J5W:A Rd.HMMs. While repacking the rotamers into the theoretical 3D-structures, Rosetta-design identifies sites of low or no variation, with higher informational content. Clearly these sites are distributed in a rather different way on the

1ATI:A, or 1J5W:A Rd.HMMs (Fig. 3 B and C), and only a fraction of these low-variance sites do coincide with structurally equivalent sites (Fig. 3A), making each Rd.HMM different enough.

A similar analysis of the lysozymes from lambda phage (or *E. coli*), T4 phage, chicken (*Gallus gallus*) and goose (*Anser anser anser*) led to very similar results.

A somehow artificial example comes from the Rosetta-designed non-natural proteins Top7 and M. This example is used to illustrate the interpretation of the Rd.HMM information in the next section.

8. The Rd-HMMer protocol: A practical guide

This section describes how to generate a Rd.HMM and interpret the results.

1. Software requirements:
   - Rosetta suite v. 2.3 or above. The examples given here apply to v. 2.3, but the porting to v. 3.1 is straightforward. Rd v. 2.1 is considerably faster, but exploration of the sequence space is better in Rd. v 3.1.
   - HMMer v. 2 or above. The examples given here were done with v. 3, which is considerably faster, therefore recommended.
   - VMD v. 1.8.7 or above. (Humphrey et al., 1996)
   - SwissPDB viewer v. 4, or above (Kaplan &Littlejohn, 2001).
   - Sequence databases. You may download the protein nr, SeqRef or UniProt-Sprot databases from the NCBI site (Sayers et al., 2010), or any other fulfilling your needs. As an alternative, you may prepare a small database using psi-blast at any server. It is recommended to include not only the sequences of proteins related to the structure of interest, but also other unrelated sequences, preferably selected at random.

VMD and SwissPDB viewer are not essential but are very useful for PDB file manipulation.

2. Prepare your PDB file. Rd v. 2.3 requires your PDB file to have non-zero beta factors. Residues may be absent, as long as none of the corresponding backbone heavy atoms are present. Therefore atoms with types C, CA, N, C and O for a particular residue should be all present for each residue, or all absent. For an incomplete residue you may open your file with Swiss PDB viewer. This program will rebuild the missing atoms, which is recommended for models; but you can use the software to completely remove the residue, which is preferable for experimental data. A special case is the oxygen atom of the C-terminus (OXT), which is required by Rd. This atom can be rebuilt with SwissPDB viewer, but this is not done automatically. An alternative to Swiss PDB viewer is VMD using the PFSgen plugin. Although PDB manipulation in VMD requires more experience, its scripting language is more powerful.

If your structural PDB file comes from a modeling exercise, review the geometrical and sterical quality of your model. If required, refine it with molecular mechanics software. A

more detailed description of this kind of problems in protein modeling and how to fix them can be found elsewhere (Chavelas Adame et al., 2011; Rosales-León et al., 2012).

3. Build many replicates of your PDB file with a random assignment of amino acid sequence. This can be done in two ways:

a. *With VMD*. Use the *atomselect* command to select the backbone atoms of all residues, one at a time, and change the residue name to any amino acid selected at random. In the C-terminal residue make sure to include the OXT atom in your selection. Then select all backbone atoms including the OXT and save the file. A script to do this with VMD can be requested to the corresponding author.

b. *With Rd*. Prepare several Rosetta input *resfile* with a tag *PICKAA X* (replace X for a random 1-letter amino acid code) for every position in the PDB file of interest. Then run Rd, once for each *resfile* you made, with the following command:

rosetta.gcc -s 1QYS.pdb-design -fixbb -chain A -resfile 1QYSaa.res-ndruns 1 -pdbout 1QYSAaa

Here, we assume 1QYS.pdb to be the starting PDB file, *A* to be the subunit of interest, *1QYSaa.res* is the *resfile*, and your result will be named *1QYSAaa_0001.pdb* (depending on the version, you will also need a paths.txt in your folder, or the path to the rosetta database should be indicated in the command line). In Rd v. 3.1, the *resfile* format and some command line options have changed (check Rosetta documentation for details).

This step can be repeated at will, to create many sequence-randomized PDB files, but in our experience, at least 10 are needed for a reliable HMM.

4. Rebuild each sequence-randomized PDB file with Rd. First you will need a *resfile* with the tag *ALLAA* (1QYSall.res), and a text file (pdb4rbld.lst) containing the names of all the sequence-randomized PDB file created in the previous step, one per line. Then, rebuild each input file 29 times using the command:

rosetta.gcc -design -fixbb -chain A-l pdb4rbld.lst -resfile 1QYSall.res -ndruns 29 -pdbout

You can generate many rebuilt PDBs per input file, but you need at least 100 sequences in the end to produce a representative HMM. In our experience, a better exploration of the sequence space results from many sequence-randomized input files and between 10 to 30 rebuilt PDBs for each input PDB file.

5. Extract the amino acid sequence for each rebuilt PDB file and save it in a text file (*1QYSA_a-O.fas*), in fasta format. This file represents an alignment, though a sequence alignment software is not necessary, due to the reasons commented at the end of section 5. Then use HMMer to prepare a hidden Markov model of your sequence alignment:

hmmbuild --informat afa 1QYSA_a-O.hmm 1QYSA_a-O.fas

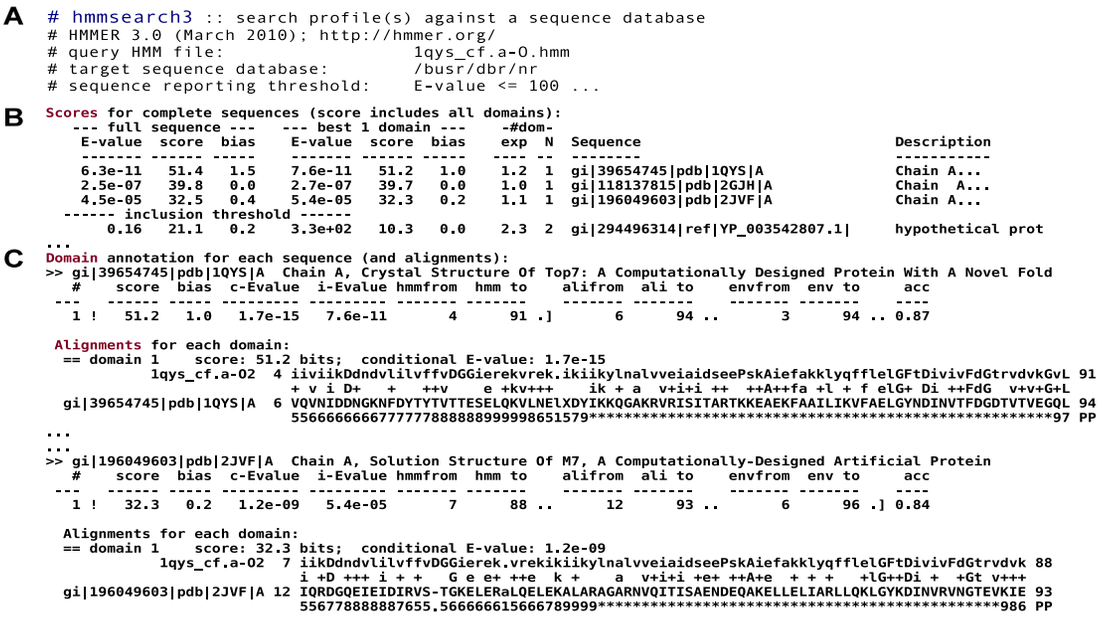
If you are using HMMer v. 2.0, you need to calibrate your model with:

hmmcalibrate 1QYSA_a-O.hmm

6. Search a sequence database (*i.e.* NCBI-nr) with:

hmmsearch -E 100 -Z 10000000 1QYSA_a-O.hmm path2db/nr

Here a local copy of the nr is assumed to be in your system in fasta format. The -Z flag will scale the E-values to 10 million sequences. This is recommended to make E-values comparable, because E-values are linearly dependent on the size of the sequence database searched. The default E value is 10, but here it was set to 100 to lower the search threshold.




```
# hmmsearch3 :: search profile(s) against a sequence database
# HMMER 3.0 (March 2010); http://hmmer.org/
# query HMM file:                  1qys_cf.a-O.hmm
# target sequence database:        /busr/dbr/nr
# sequence reporting threshold:    E-value <= 100 ...

Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence                          Description
    ------- ------ -----    ------- ------ -----   ---- --  --------                          -----------
    6.3e-11   51.4   1.5    7.6e-11   51.2   1.0    1.2  1  gi|39654745|pdb|1QYS|A            Chain A...
    2.5e-07   39.8   0.0    2.7e-07   39.7   0.0    1.0  1  gi|118137815|pdb|2GJH|A           Chain  A...
    4.5e-05   32.5   0.4    5.4e-05   32.3   0.2    1.1  1  gi|196049603|pdb|2JVF|A           Chain A...
  ------ inclusion threshold ------
       0.16   21.1   0.2    3.3e+02   10.3   0.0    2.3  2  gi|294496314|ref|YP_003542807.1|  hypothetical prot
...
Domain annotation for each sequence (and alignments):
>> gi|39654745|pdb|1QYS|A  Chain A, Crystal Structure Of Top7: A Computationally Designed Protein With A Novel Fold
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !   51.2   1.0   1.7e-15   7.6e-11       4      91 .]       6      94 ..       3      94 .. 0.87

 Alignments for each domain:
  == domain 1    score: 51.2 bits;  conditional E-value: 1.7e-15
             1qys_cf.a-O2  4 iiviikDdndvlilvffvDGGierekvrek.ikiikylnalvveiaidseePskAiefakklyqfflelGFtDivivFdGtrvdvkGvL 91
                             + v i D+   +   ++v    e +kv+++    ik + a  v+i+i ++  ++A++fa +l + f elG+ Di ++FdG  v+v+G+L
  gi|39654745|pdb|1QYS|A  6 VQVNIDDNGKNFDYTYTVTTESELQKVLNElXDYIKKQGAKRVRISITARTKKEAEKFAAILIKVFAELGYNDINVTFDGDTVTVEGQL 94
                            5566666666777777888888999998651579*****************************************************97 PP
...
...
>> gi|196049603|pdb|2JVF|A  Chain A, Solution Structure Of M7, A Computationally-Designed Artificial Protein
   #    score  bias  c-Evalue  i-Evalue hmmfrom  hmm to    alifrom  ali to    envfrom  env to     acc
 ---   ------ ----- --------- --------- ------- -------    ------- -------    ------- -------    ----
   1 !   32.3   0.2   1.2e-09   5.4e-05       7      88 ..      12      93 ..       6      96 .] 0.84

  Alignments for each domain:
  == domain 1    score: 32.3 bits;  conditional E-value: 1.2e-09
              1qys_cf.a-O2  7 iikDdndvlilvffvDGGierek.vrekikiikylnalvveiaidseePskAiefakklyqfflelGFtDivivFdGtrvdvk 88
                              i +D +++ i + +   G e e+ ++e  k +    a  v+i+i +e+ ++A+e  + + +   +lG++Di +  +Gt v+++
  gi|196049603|pdb|2JVF|A 12 IQRDGQEIEIDIRVS-TGKELERaLQELEKALARAGARNVQITISAENDEQAKELLELIARLLQKLGYKDINVRVNGTEVKIE 93
                             556778888887655.566666615666789999*********************************************986 PP
```

Figure 4. An HMMer search output result. The search was done using an Rd.HMM from Top7 (PDB id. 1QYS) and the NCBI-nr as database. (A) heading, (B) scores, (C) domain-parsed scores and alignments. The statistics at the end and some information was removed for brevity. The format is as in HMMer 3.0.

An extract of the results from a typical search is presented in figure 4. The HMMer search output will report three sections: (a) Heading, (b) scores for complete sequences, (c) domain parsing, alignments and statistics. As it can be seen, according to the information in the Rd.HMM from Top7, the Top7 amino acid sequence fits into the Top7 3D-atomic coordinates (1QYS). The most relevant sections are the scores and the alignment sections. Notice how this X-ray solved 3D-stucture reports an HMM score of 51.2, matching the sequence from amino acid 3 to 94, that gives a ratio of 0.56, close to the 0.6 average value for X-ray solved structures. The reason behind the relationship is not simple, but it holds for most X-ray solved structures (with a few exceptions) (Martínez-Castilla & Rodríguez-Sotres 2010). The second hit is the C-terminal fragment of Top7 solved by NMR, the score is 39.8 for a fragment of length 50 (ratio = 0.79). This last score is higher than the score for the complete sequence, because as shown in figure 4, the C-terminus has higher proportion of local coincidences to the HMM. In the alignment to the full sequence, the contribution of the N-terminus lowers the overall score. The alignment for the C-terminal fragment was omitted because it is identical to the 1QYS alignment from position 44 to 91 (Fig. 4C). The alignment shows a consensus for the hidden Markov model, as a reference, then the sequence found aligned separated by an intermediate mask. Uppercase letters indicate

strong conservation, lower case letter conservative changes and plus sign a positive local score. The lower line, absent in HMMer 2 is the encoded posterior probability (d=0...9,*; * equals 9.5), where the approximate value of posterior probability for each site is given by equation (1).

$$pp = d \times 0.1 \; + \; 0.025 \tag{1}$$

The final hit in the search in figure 4 is the M artificial protein. This protein was designed with Rosetta-design using the same Top7 folding. Its sequence is different, but it belongs to the same family of Rd proteins. The score is smaller than for Top7 (ratio of 32/(88-7), or 0.395), but still above 0.3 and with high statistical significance. Although Rd was used to design these proteins, the concordance reveals the robustness of the amino acid assignment made by Rd, and gives further support to the structurally aware nature of the Rd.HMM alignments.

The alignment is very useful to protein modeling, because it reveals the distribution of coincident regions between the 3D-atomic coordinates of the backbone and the amino acid sequence in the database. The following features are to be taken into account:

a. Frame shift. If the residue number in the 3D-structure has an offset relative to the amino acid numbering in the sequence, either from the beginning, or starting at some intermediate site; this is usually a sign of a wrong threading of the model and the template during the modeling step. In the example, there is a difference in amino acid numbering, but this is not a frame shift, as the first residue solved in the PDB is ASP-3, corresponding to node one in the HMM, then the first 3 HMM nodes did not match the Top7 sequence and were discarded by HMMer search making the first match to residue 6, at HMM node 4.
b. Insertion/deletions. An insertion in the sequence appears as a dot in the HMMer consensus, a deletion as dashes in the sequence found. Such changes are expected if the sequence is a homologue, and not the natural sequence that corresponds to the 3D-structures analyzed with Rd.HMM. They may occur also when the PDB file has some missing amino acids (this happens frequently, due to experimental limitations of X-ray crystalography). If so, you expect this insertions to match the missing amino acids. For *in silico* modeled structures this means a local threading error, or a local defect in the model.
c. Distribution of conserved sites. The higher the number of conserved sites, the better the model. However, some strained conformations have lower energy for glycine, proline and asparagine than for every other amino acid and these residues tend to appear as strongly conserved (Uppercase letters in the mask line, and in the Rd.HMM consensus). If the sequence conservation observed is dominated by these residues, you model may be wrong, even if your score has statistical significance.

9. Guiding the 3D-modeling of proteins with Rd-HMMer

There are many publications describing different approaches to the solution to the protein folding problem (Roy et al., 2010; Kim et al., 2004; Karplus, 2009; Melo & Feytmans, 1998) but most of them focus on the theory, or present a technical treatment. Fisher and Sali

published a practical guide to the use of the popular modeling software MODELLER (Fiser & Sali 2003), where many useful hints are given. Recently, Chavelas-Adame and coworkers published a guide with emphasis on the use of open software [45]. The present account will not attempt to repeat the work, and only the most important conclusions are given here:

a. Many servers and software programs are available to aid the comparative modeling of proteins (Roy et al., 2010; Kim et al., 2004; Karplus, 2009; Melo & Feytmans, 1998; Rosales-León et al., 2012; Fiser & Sali 2003), some options are available for *ab initio* modeling (Kryshtafovych et al., 2009; Kim et al., 2004; Srinivasan et al., 2004; Xu & Zhang, 2012), and this list is far form complete. None of them achieves 100% success, and even the most successful can fail where other, usually less reliable, may succeed (Kryshtafovych et al., 2009; Melo & Feytmans, 1998).
b. A model is fundamentally wrong when the folding pattern in the model bears little or no relationship to the true native fold. Some models may offer a good approximation, but have wrong geometrical, sterical and/or chemical features at some locations, *i.e.* the bond lengths, angles, sidechain-sidechain contact distances and orientation may have important deviations from the expected values found in known chemical structures. This last kind are usually designated as unrefined models.
c. Unrefined models can be recognized with various energy scoring strategies (Luthy et al., 1992; Shi et al., 2009; Hu & Jiang, 2010; Melo & Feytmans, 1998); and can be corrected through the use of molecular mechanics software (Rosales-León et al., 2012; Fiser & Sali 2003), though this approach has limitations, as mentioned before (Faver et al., 2011; Hu & Jiang, 2010; Melo & Feytmans, 19985).
d. Wrong models instead may frequently be deceitful, because, due to their systematic error [5], a molecular mechanics force-field may report a low energy value, as long as the chemical and geometrical details are well refined. Rd.HMM offers a solution to this problem, because these models will produce an HMM search report with no hits, or will score sequences, other than the modeling target (Chavelas Adame et al., 2011; Rosales-León et al., 2012).
e. The analysis of the Rd.HMM search report may help in the identification of errors in the alignment between the target amino acid sequence and the template selected for comparative modeling. If you find a frame-shift or an unexpected insertion/deletion pair, you can use the HMM search alignment and realign the target sequence and the template. MODELLER is a very good choice for that aim (Fiser & Sali 2003). In addition, a wrongly threaded model can be recycled by replacing the consensus sequence with the PDB sequence in the model (which is the target sequence), and producing a target to target alignment, with the insertions and deletions suggested by HMMer. MODELLER can then be used to generate new models. This last procedure is only recommended if your HMMer score is positive and has good statistical significance, for otherwise, the structural inaccuracy of the Rd.HMMs becomes a serious issue.
f. Comparative modeling has been extended thanks to methods able to find templates with low sequence homology to the target (Wallace et al., 2005; Karplus, 2009). But sometimes the selected template is too distant. If the Rd.HMM of the candidate structures are obtained, these can be use to score the target sequence. The resulting

scores, statistical significance and the alignment may guide your template selection. However, if the Rd.HMM of a template candidate gives a negative score, and still you decide to use it, do not trust the Rd.HMM alignment without further improvement using other tools, as it may be seriously flawed.

g. Finally, if you use the ROSETTA suite or the ROBETTA server to produce your models, these structures are expected to have a ROSETTA-like bias, *i.e.* their Rd.HMM scores will increase and a good model with this bias is expected to have a ratio of HMMer score to sequence length close to one. While in models produced with other software a Rd.HMM score ratio of 0.3 is acceptable, in a ROSETTA produced model this score is low and may reflect important flaws. Look at the alignment carefully, as recommended in the previous section.

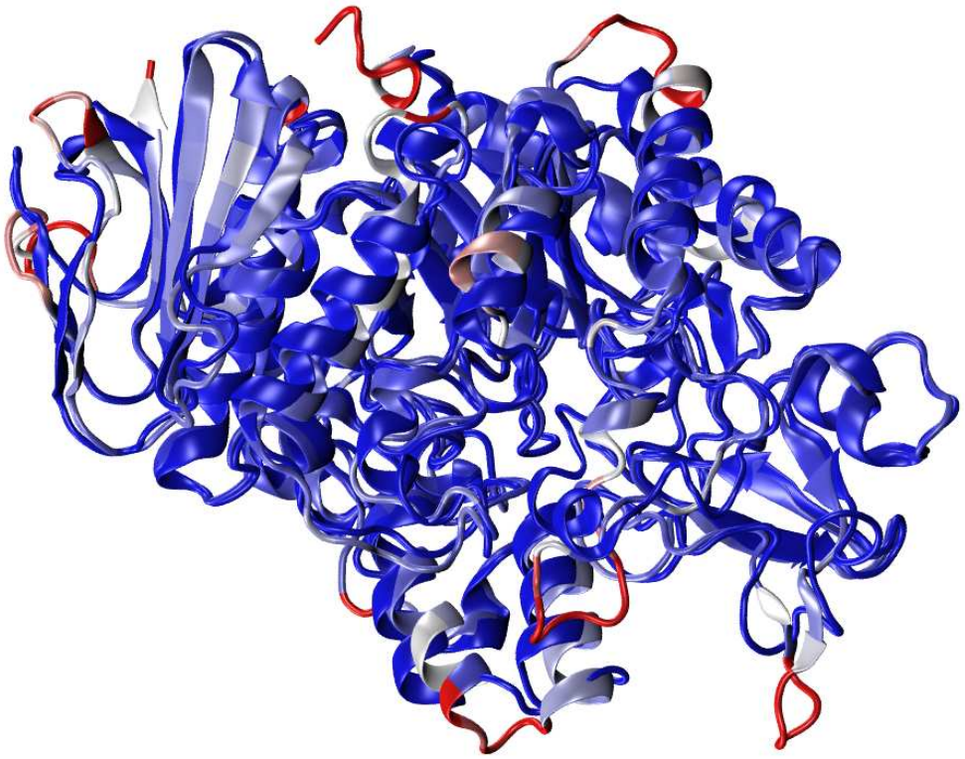


Figure 5. Comparison of the yeast α-glucosidase model produced included in the publication by Brindis et al (Brindis et al., 2011), with the X-ray solved structure of its homologue, the yeast isomaltase (Yamamoto et al., 2010). The isomaltase is shown as blue cartoons and the α-glucosidase cartoons are colored according to the amino acid rmsd from isomaltase, ranging from very low (blue) to intermediate (white) to high (red).

As an example of the advantages of Rd.HMM, we refer to two cases of recent success. Brindis and coworkers (Brindis et al., 2011) analyzed the effects of a natural product on α-glucosidase. This work reports a model for the budding yeast α-glucosidase used to analyze the molecular grounds for the (Z)-3-butylidenephthalide inhibitory action. In the preparation of the model, Rd.HMM allowed to detect a threading problem (insertion/deletion pair) in one β-strand in the core of the model. While the sheet was slid only a few Å from its position, the contact with neighboring strands completely distorted

the chemical interaction network affecting the model stability. The correction of this problem and the use of molecular dynamics simulations led to a well refined and reliable model with a good Rd.HMM score. A few months later (when the paper was in press) the 3D-structure of a close homologue (isomaltase) was released (Yamamoto et al., 2010). The X-ray data corroborated the model quality, as the model core backbone has an rmsd of 1.81 Å form the experimental data. Figure 5 shows a superposition of both structures colored by backbone rmsd form blue (low) to white (medium) to red (high).

In a second example, the 3D-structure of two isoforms of plant inorganic pyrophophatases was obtained using a combined strategy of web servers, MODELLER and molecular dynamics simulations. The resulting models provided ground for the lack of quaternary structure in plant pyrophosphatases (Rosales-León et al., 2012). Although the sequences of several related isoforms were initially sent to the servers, only one isoform was correctly modeled, according to Rd.HMM, but the Rd.HMM of the good model gave an alignment for the sequence of a second isozyme. This alignment, and the correct model were then used to produce the second model. Though this last model was not directly based on experimental data, its quality was high, according to Rd.HMM (Rosales-León et al., 2012).

10. Rd-HMMer limitations

Since most Rd.HMM limitations have been mentioned. We only summarize them here:

a. Rd.HMM sensitivity makes it useful for medium to good quality models. Low quality models, may still be of use as starting points, but the Rd.HMM data will only indicate the low quality and will not allow to discriminate a wrong model from an unrefined one.
b. The structurally aware nature of the Rd.HMM alignments is to be trusted only for good quality models. As the Rd.HMM score drops, the sequence to structure correlation becomes weak.
c. Rd.HMM does not offer much information on how to modify the model to improve its appropriateness, other than the presence of insertion/deletions or sequence to structure frame-shifts.
d. A model may be badly refined and get a good Rd.HMM score, as long as Rosetta-design is able to process the backbone coordinates and repack the residues. Therefore, the Rd.HMM score is insufficient information. Information from other software, such as ANOLEA energy (Melo & Feytmans, 1998) or molecular mechanics energy (Hu & Jiang, 2010) is always required to test a model quality.
e. Finally, there is no formal proof for the perfect correspondence between a Rd.HMM high score and the prediction for the 3D-structure of a protein to be native-like. Therefore, from two predictions, of which only one represents the native fold, it might be possible to produce a high Rd.HMM score for the target sequence (a false positive). However, despite our best efforts we have only found the false negative case, *i.e.* a good prediction (or even a 3D-structure from experimental data) may give a low Rd.HMM score. To the best of our knowledge, among the quality assessment methods, this feature is unique to the Rd.HMM protocol.

11. Conclusions and perspectives

Although the Rd.HMM protocol is highly sensitive and its alignments become inaccurate when the HMM score decreases, it can be used to guide the comparative modeling of proteins, as the examples given in section 8 show. Even if the alignment employed is flawed, when the model is produced and analyzed with Rd.HMM, the flaw will become evident and the model can then be discarded, and additional modeling rounds may be tried.

An additional advantage of Rd.HMM alignments, as a guide to comparative modeling, comes form the fact that Rd.HMM models are independent of the functional constrains reflected in the conservation of active and binding sites. Since the Rd. step removes all conservation due to ligand binding and functional sites, other than that required to keep the structure stable, geometrical differences in the organization of two related, but not identical active sites will not affect the modeling process. In contrast, in the classic comparative modeling methods, the residue conservation at active and other functional sites is usually an important reference to perform the sequence to structure alignment. Then when a model in produced with the guidance of Rd.HMM, and a model with good quality and appropriateness is obtained, any coincidences in the active site geometry, would not come as a consequence of forcing the conserved residues in the target sequence to fall at the template's active site, but should be a consequence of meeting the structural requirements of the target.

From the above discussion, Rd.HMM is clearly a valuable tool, but has some limitations. We speculate that some of this limitations derive form the inability of HMMs to incorporate long range interactions, which can be detected as significant mutual information between distant positions in the sequence alignments. Currently we are working on the analysis of the mutual information in the Rosetta-designed sequence alignments using the statistical coupling analysis strategy (Socolich et al., 2005,Lockless et al., 1999). We hope this powerful statistical approach can extend the Rd.HMM and provide a richer tool.

Author details

León P. Martínez-Castilla and Rogelio Rodríguez-Sotres
Facultad de Química, Universidad Nacional Autónoma de México, México

Acknowledgement

Funding PAPIIT-DGAPA-UNAM IN210212, CONACyT CB2008-1-101186, PAIP-FQ-UNAM 4290-09, PAIP-FQ-UNAM 4290-07.

Abbreviations

NP-hard problem, as hard to solve as an NP-complete problem; NP-complete problem, no algorithm taking a polynomial-time exists for its solution; Rd, ROSETTA-design; MM,

Markov model; HMM, hidden Markov model; CASP, critical assessment of the structure of proteins; PDB, international protein data bank; TIM, triose phosphate isomerase.

12. References

Bowie, J., Reidhaar-Olson, J., Lim, W., & Sauer, R.(1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science,* Vol. 247, No. 4948, (Mar 1990) pp. 1306-1310, ISSN 0036-8075 (print), 1095-9203 (electronic)

Brindis, F., Rodríguez, R., Bye, R., González-Andrade, M., & Mata, R.(2011). (Z)-3-butylidenephthalide from *Ligusticum porteri* , an α-glucosidase inhibitor. *J Nat Prod,* Vol. 74, No. 3, (Sep 2011) pp. 314-20, ISSN 0163-3864 (print) 1520-6025 (electronic)

Butterfoss, G. L. & Kuhlman, B.(2006). Computer-based design of novel protein structures. *Annu Rev Biophys Biomol Struct,* Vol. 35, (Jun 2006) pp. 49-65, ISSN 1056-8700

Chavelas Adame, E. A., Hernández-Domínguez, E. E., Gaytán-Mondrangón, S., Rosales León, L., Valencia-Turcotte, L., & Rodríguez-Sotres, R.(2011). A Hitchhiker's Guide to the modeling of the three-Dimensional structure of proteins. *International Color Biotechnology Journal,* Vol. 1, No. 1, (Nov 2011) pp. 26-35, ISSN 2226-0404 (electronic)

Cheng, G., Qian, B., Samudrala, R., & Baker, D.(2005). Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Research,* Vol. 33, No. 18, (Sep 2005) pp. 5861--7, ISSN 1362-4962 (print)

Chivian, D. & Baker, D.(2006). Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Research,* Vol. 34, No. 17, (Sep 2006) pp. e112, ISSN 1362-4962

Chodanowski, P., Grosdidier, A., Feytmans, E., & Michielin, O.(2008). Local Alignment Refinement Using Structural Assessment. *PLoS ONE,* Vol. 3, No. 7, (Jul 2008) pp. e2645, ISSN 1932-6203 (electronic)

Cowles, M. K. & Carlin, B. P.(1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association,* Vol. 91, No. 434, (Jun 1996) pp. 883-904, ISSN 0162-1459 (print), 1537-274X (electronic)

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R.(2008). The protein folding problem. *Annu Rev Biophys,* Vol. 37, (Jun 2008) pp. 289-316, ISSN ISSN:1936-122X (print) 1936-1238 (electronic) 1936-122X (linking)

Eddy, S. R., Mitchison, G., & Durbin, R.(1995). Maximum discrimination hidden Markov models of sequence consensus. *J Comput Biol,* Vol. 2, No. 1, (Jan 1995) pp. 9--23, ISSN 1066-5277 (print); 1557-8666 (electronic)

Eddy, S. R.(2004). What is a hidden Markov model? *Nat Biotech,* Vol. 22, No. 10, (Oct 2004) pp. 1315--1316, ISSN 1087-0156

Faver, J. C., Benson, M. L., He, X., Roberts, B. P., Wang, B., Marshall, M. S., Sherrill, C. D., & Merz, Jr., K. M.(2011). The Energy Computation Paradox and *ab initio* Protein Folding. *PLoS ONE,* Vol. 6, No. 4, (Apr 2011) pp. e18868, ISSN e1932-6203

Fiser, A. & Sali, A.(2003). Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol,* Vol. 374, (Dec 2003) pp. 461-91, ISSN 978-0-12-182777-9

Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., Boomsma, W., Bottaro, S., & Ferkinghoff-Borg, J.(2010). Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized. *PLoS ONE,* Vol. 5, No. 11, (Nov 2010) pp. e13714, ISSN e1932-6203

Hart, W. E. & Istrail, S.(1997). Robust proofs of NP-hardness for protein folding: general lattices and energy potentials. *J Comput Biol,* Vol. 4, No. 1, (Jan 1997) pp. 1-22, ISSN 1066-5277 (print); 1557-8666 (electronic)

He, X. & Merz, K. M.(2010). Divide and Conquer Hartree-Fock Calculations on Proteins. *Journal of Chemical Theory and Computation,* Vol. 6, No. 2, (Jan 2010) pp. 405-411, ISSN 1549-9618 (print) 1549-9626 (electronic)

Hu, Z. & Jiang, J.(2010). Assessment of biomolecular force fields for molecular dynamics simulations in a protein crystal. *J Comput Chem,* Vol. 31, No. 2, (Jan 2010) pp. 371-80, ISSN 1096-987X

Humphrey, W., Dalke, A., & Schulten, K.(1996). VMD: visual molecular dynamics. *J Mol Graph,* Vol. 14, No. 1, (Feb 1996) pp. 33-38, ISSN 1093-3263

Ilyin, V. A., Abyzov, A., & Leslin, C. M.(2004). Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Science: A Publication of the Protein Society,* Vol. 13, No. 7, (July 2004) pp. 1865-1874, ISSN 0961-8368

Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., & Baker, D.(2008). *De novo* computational design of retro-aldol enzymes. *Science (New York, N.Y.),* Vol. 319, No. 5868, (Mar 2008) pp. 1387-1391, ISSN 0036-8075 (print), 1095-9203 (electronic)

Kaplan, W. & Littlejohn, T. G.(2001). Swiss-PDB Viewer (Deep View). *Briefings in Bioinformatics,* Vol. 2, No. 2, (May 2001) pp. 195-197, ISSN 1477-4054 (print) 1467-5463 (electronic)

Karplus, K.(2009). SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res,* Vol. 37, No. Web Server issue, (July 2009) pp. W492-7, ISSN 1362-4962 (print)

Kim, D. E., Chivian, D., & Baker, D.(2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res,* Vol. 32, No. Web Server issue, (Jul 2004) pp. W526-W531, ISSN 1362-4962 (print) 0305-1048 (electronic)

Kono, H. & Saven, J. G.(2001). Statistical theory for protein combinatorial libraries. packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *Journal of Molecular Biology,* Vol. 306, No. 3, (Feb 2001) pp. 607 - 628, ISSN 0022-2836

Kryshtafovych, A., Krysko, O., Daniluk, P., Dmytriv, Z., & Fidelis, K.(2009). Protein structure prediction center in CASP8. *Proteins,* Vol. 77, No. Suppl 9, (July 2009) pp. 5-9, ISSN 1097-0134

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D.(2003). Design of a novel globular protein fold with atomic-level accuracy. *Science,* Vol. 302, No. 5649, (Nov 2003) pp. 1364--1368, ISSN 0036-8075 (print), 1095-9203 (electronic)

Lathrop, R. H.(1994). The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Engineering,* Vol. 7, No. 9, (Sep 1994) pp. 1059-1068, ISSN 1741-0134 (print) 1741-0126 (electronic)

Levinthal, C.(1968). Are there pathways for protein folding? *Journal de Chimie Physique et de Physicochimie Biologique,* Vol. 65, No. 1-4, (Jan 1968) pp. 44-45, ISSN 0021-7689

Lockless, S. W. & Ranganathan, R.(1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science,* Vol. 286, No. 5438, (Oct 1999) pp. 295-299, ISSN 0036-8075 (print), 1095-9203 (electronic)

Luthy, R., Bowie, J. U., & Eisenberg, D.(1992). Assessment of protein models with three-dimensional profiles. *Nature,* Vol. 356, No. 6364, (Mar 1992) pp. 83--85, ISSN 0028-0836 (print)

Martínez-Castilla, L. P. & Rodríguez-Sotres, R.(2010). A score of the ability of a three-dimensional protein model to retrieve its own sequence as a quantitative measure of its quality and appropriateness. *PLoS One,* Vol. 5, No. 9, (Sep 2010) pp. e12483, ISSN e1932-6203

Melo, F. & Feytmans, E.(1998). Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol,* Vol. 277, No. 5, (Apr 1998) pp. 1141-1152, ISSN 0022-2836

Mengersen, K. L. & Tweedie, R. L.(1966). Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics,* Vol. 24, No. 1, (Feb 1966) pp. 101-121, ISSN 0090-5364

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E.(1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics,* Vol. 21, No. 6, (Jun 1953) pp. 1087-1092, ISSN 0021-9606 (print), 1089-7690 (electronic)

Pawlowski, M., Gajda, M. J., Matlak, R., & Bujnicki, J. M.(2008). MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformatics,* Vol. 9, No. Sep, (Sep 2008) pp. 403, ISSN 1471-2105

Raha, K. & Merz, Jr, K. M.(2005). Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J Med Chem,* Vol. 48, No. 14, (Jul 2005) pp. 4558-4575, ISSN 0022-2623 (print) 1520-4804 (electronic)

Rodriguez, R., Chinea, G., Lopez, N., Pons, T., & Vriend, G.(1998). Homology modeling, model and software evaluation: three related resources. *Bioinformatics,* Vol. 14, No. 6, (Jul 1998) pp. 523-528, ISSN 1460-2059 (print) 1367-4803 (electronic)

Rosales-León, L., Hernández-Domínguez, E. E., Gaytán-Mondragón, S., & Rodríguez-Sotres, R.(2012). Metal binding sites in plant soluble inorganic pyrophosphatases. An example of the use of ROSETTA design and hidden Markov models to guide the homology modeling of proteins. *Journal of the Mexican Chemical Society,* Vol. 56, No. 1, (Jan-Mar 2012) pp. 23-31, ISSN 1665-9686

Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M., & Bourne, P. E.(2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res,* Vol. 39, No. Database issue, (Jan 2011) pp. D392-D401, ISSN 1362-4962 (print) 0305-1048 (electronic)

Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., & Baker, D.(2008). Kemp elimination catalysts by computational enzyme design.

Nature, Vol. 453, No. 7192, (May 2008) pp. 190-195, ISSN 0028-0836 (print) 1476-4687 (electronic)

Roy, A., Kucukural, A., & Zhang, Y.(2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc,* Vol. 5, No. 4, (Mar 2010) pp. 725-738, ISSN 1754-2189 (print) 1750-2799 (electronic)

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., John Wilbur, W., Yaschenko, E., & Ye, J.(2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* Vol. 38, No. suppl 1, (Jan 2010) pp. D5-16, ISSN 1362-4962 (print) 0305-1048 (electronic)

Schuster-Böckler, B. & Bateman, A.(2005). Visualizing profile, profile alignment: pairwise HMM logos. *Bioinformatics,* Vol. 21, No. 12, (Jun 2005) pp. 2912-2913, ISSN 1367-4803 (print) 1460-2059 (electronic)

Shapiro, L. W. & Zeilberger, D.(1982). A Markov chain occurring in enzyme kinetics. *Journal of Mathematical Biology,* Vol. 15, No. 3, (Nov 1982) pp. 351-357, ISSN 0303-6812

Shi, S., Pei, J., Sadreyev, R. I., Kinch, L. N., Majumdar, I., Tong, J., Cheng, H., Kim, B.-H., & Grishin, N. V.(2009). Analysis of CASP8 targets, predictions and assessment methods. *Database (Oxford),* Vol. 2009, (Apr 2009) pp. bap003, ISSN 1758-0463

Slovic, A. M., Kono, H., Lear, J. D., Saven, J. G., & DeGrado, W. F.(2004). Computational design of water-soluble analogues of the potassium channel KcsA. *Proc Natl Acad Sci U S A,* Vol. 101, No. 7, (Feb 2004) pp. 1828-1833, ISSN 1091-6490

Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., & Ranganathan, R.(2005). Evolutionary information for specifying a protein fold. *Nature,* Vol. 437, No. 7058, (Sep 2005) pp. 512-8, ISSN 0028-0836 (print) 1476-4687 (electronic)

Srinivasan, R., Fleming, P. J., & Rose, G. D.(2004). *Ab initio* protein folding using LINUS. *Methods Enzymol,* Vol. 383, (Apr 2004) pp. 48-66, ISSN 978-0-12-391860-4

Wallace, I. M., Blackshields, G., & Higgins, D. G.(2005). Multiple sequence alignments. *Curr Opin Struct Biol,* Vol. 15, No. 3, (Jun 2005) pp. 261-266, ISSN 0959-440X

Wallner, B. & Elofsson, A.(2003). Can correct protein models be identified? *Protein Sci,* Vol. 12, No. 5, (May 2003) pp. 1073-1086, ISSN 1469-896X

Wilmanns, M. & Eisenberg, D.(1995). Inverse protein folding by the residue pair preference profile method: estimating the correctness of alignments of structurally compatible sequences. *Protein Eng,* Vol. 8, No. 7, (July 1995) pp. 627-39, ISSN 1741-0134 (print) 1741-0126 (electronic)

Xu, D., Zhang, J., Roy, A., & Zhang, Y.(2011). Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based *ab initio* folding and FG-MD-based structure refinement. *Proteins,* Vol. 79, No. S10, (Aug 2011) pp. 147-160, ISSN 1097-0134

Yamamoto, K., Miyake, H., Kusunoki, M., & Osaki, S.(2010). Crystal structures of isomaltase from *Saccharomyces cerevisiae* and in complex with its competitive inhibitor maltose. *FEBS J,* Vol. 277, No. 20, (Oct 2010) pp. 4205-14, ISSN 1742-4658