Alexander Örnbratt

Review of the Topic *Co-evolution* in Structural Bioinformatics

Assignment 4 - TDA507

December 19, 2019

1 What is *co-evolution*?

Evolution is, of course, a fundamentally important and captivating area of study within many fields of science in the modern world, teaching mankind about how organisms of the past may have become what they are today, as well as presenting possible ideas about what the future may hold for life as we know it. On a microscopic level, mutations happen frequently and rapidly, and it is these mutations that eventually give rise to the evolution of a species. Understanding the mutations and evolution of a protein structure or sequence is absolutely integral for comprehending the way that that protein will behave and function. Predicting these protein structures computationally, the central focus of bioinformatics, would increase scientists' knowledge on the subject at a pace not possible experimentally [1, p. 2].

Co-evolution refers to different amino acid residues interacting in such a way that they have mutual influence on their respective mutations and evolution. There may exist pairs of amino acids within a protein that regularly evolve together, and may therefore be related in terms of distance. Identifying these pairs of residues can lead to a better understanding of the structure of the protein, and consequently how it functions, thus presenting great benefits in the field of, for example, therapeutic drug creation.

The computational problem being addressed when discussing the topic of co-evolution is that given only a protein sequence, it was thought to be near-impossible to predict how that protein would fold when created, and thus to know its structure. Scientists of course know the structures of many thousands of proteins to a very high degree of accuracy, but they have been determined by experimental methods such as X-ray crystallography or NMR spectroscopy, which take a long time to perform, making it infeasible to find the structures of all known proteins in this way. The basic idea of co-evolution-related problem solving is to use evolutionary data about protein sequences, along with a model for interpreting that data, in order to lessen the number of 3D conformations that that protein could take when folded [1, p. 3].

2 Computational Methods and Approaches

To implement a solution to the problem, researchers must try to come up with computer-based methods that can try to, given a number of inputs, predict the three-dimensional atomic structure of a protein. Recent advances in the field of structural bioinformatics have brought forth some promising, though not perfect, examples of methods for solving this kind of problem.

The basic process of computing a prediction of the 3D structure of a protein from only its sequence is described by Marks et al. [1, p. 14] as a "pipeline" containing a number of steps, the first being to compile a large number of sequences that are related, but "evolutionarily diverged", i.e. that these protein sequences have for example a common ancestor. These sequences must then be aligned so that the differences between their residues can more easily be found. The next step is to calculate the "covariance" for each pair of amino acids. This is essentially the strength of the correlation between changes in each of them. From these values or matrices, one can find so-called "couplings" between pairs of residues using a statistical model of choice. These couplings represent co-evolving amino acids, of which, the highest scoring ones are theorized to be in close proximity to one another in the three-dimensional structure of the protein, and are called "contacts". These predicted contacts can then be used as constraints when simulating the folding of the protein, which is done by the method of for example simulated annealing. What results from these simulations are a number of possible structures that the protein could have, and these are then ranked using various criteria. It is these ranked results that represent the end-product of this method and the "answer" to the computational problem. [1, p. 14]

To test this process, researchers start with simple proteins with a single domain to see if the correlation information about the amino acid pairs actually matches where the folds should theoretically be, eventually advancing to large, complex, diverse proteins if the method works well [1, pp. 5-6].

To understand the concept of "covariance" and the correlation between a pair of residues in an alignment of sequences, one must realize that the strength of this correlation is dependent on a number of factors. Horner et al. mentions that there are five factors: (1) the "phylogenetic relationships between [evolutionarily related] homologous sequences" [2, p. 47], (2) structural and (3) functional constraints, (4) the interaction between the first three factors, and (5) the randomness from sequence sampling and random co-variation [2, pp. 47-48]. Methods and algorithms for solving these types of problems most often focus on finding the constraints (aforementioned points 2 and 3) that make it possible to find folds [2, p. 47].

Regarding the statistical models mentioned in the process description above, there are a great number of these types of models and they have exhibited various degrees of success. These models or methods can aid in finding pairs of residues within proteins that closely interact, in the best cases making it possible to accurately determine how the protein folds - the most difficult thing to predict about its structure [1, p. 2] and a central part of understanding its function [1, p. 1]. Horner et al. [2, p. 53] mentions the following methods for intramolecular contact prediction: Pearson-pairwise, Tree-correlation, OMES, SCA, ELSC, Dependency, MI (mutual information), P2P, Singer, and Fariselli. In addition to SCA and MI ("local models"), Marks et al. describes the Bayesian network model (BNM) and the direct information (DI) model (categorizing them as "global models") [1, p. 10].

The methods based on the Pearson correlation coefficient work by finding a number for every pair of sites in the protein, based on comparisons of weighted substitutions between sequences [2, p. 47]. The methods based on information theory do not look at substitutions, instead focusing on the frequencies and probabilities of finding certain amino acids at pairs of sites [2, p. 48]. The MI method, an example of this, tries to find out how much knowledge that information of one site gives about another site [2, p. 48]. Another class of methods are those that calculate a statistic for every pair of residues, taking into account the difference between the observed and expected number of occurrences of amino acid pairs [2, pp. 49-50]. SCA and ELSC are examples of alignment perturbation methods, which entail comparing entire alignments with altered, or perturbed, sub-alignments to calculate the probability of a residue correlation at a specific site [2, p. 50]. There are also methods that focus primarily on finding probabilities of contacts from phylogenetic trees, methods that use log-odds substitution matrices based on existing data to predict contacts, and methods that use artificial intelligence, machine learning, and specifically neural networks to find patterns related to identifying structural contacts in the proteins [2, pp. 50-51].

One of the main disadvantages plaguing most of the above methods is that they depend on a "large number of evolutionarily related sequences" [3, p. 12]. The dependencies between different residues in proteins will not become clear if there are only a few different sequences to compare. As clearly shown in Table 2 of Horner et al. [2, p. 53], the accuracies of the various algorithms are far from perfect; no single algorithm results in a percentage greater than 30%

for the *number of true positives* divided by the *number of predictions made*, or in other words, only a fraction of the predictions made about intramolecular amino acid structural contacts were true. False positives for the distance constraints constitute one of the biggest obstacles in creating an accurate structural model of the proteins. These false positive constraints need to be avoided through the use of equations that will filter out residue coupling background noise [3, p. 15], [1, p. 15].

Marks et al. argues that global models (BNM and DI) are better than their local counterparts (e.g. MI), as they can account for and remove indirect correlations between different pairs of amino acids [1, p. 3], the significant problem in finding actual contacts being that the evolution of one amino acid may affect the evolution of several others, but obviously not *all* these amino acids would necessarily be close in three-dimensional space, so the "true" coupled pairs must be found [1, pp. 1-2]. This so-called "transitivity effect" limits the effectiveness of local methods [1, p. 3]. The major technical difference between the equations used for DI and those for MI is that MI only accounts for the observed local frequencies of amino acid pairs while DI includes an expression for the probability of doubly-constrained pairs of residues, leading to a much higher degree of accuracy. The DI algorithm will result in a more accurate and even distribution of residue contact pairs on a distance map than MI is capable of [1, p. 6].

The positive aspect of using the described process and methods to compute the 3D structure of a protein, as Marks et al. claims to be true for their DI algorithm [1, p. 16], is that they do not have to be run on a supercomputer and can in fact be completed in minutes on a regular personal computer, opening up the possibility to make these kinds of structure predictions on a large scale, for potentially limitless numbers of interesting proteins.

3 Conclusion

Fully understanding the correlations between substitutions occurring in residue pairs gives rise to a better comprehension of the evolution of proteins and subsequently the development of their structures. All three research papers cited have determined that it is possible and *practical* to use the described process to deduce the 3D structure of proteins, although each uses slightly different statistical models to get there. In conclusion, determining algorithmically where the amino acid contacts take place means that scientists can make more accurate models of the structure of proteins, speeding up the research process and making the proteins more readily available for the medical field.

References

[1] D. Marks et al., "Protein 3D Structure Computed from Evolutionary Sequence Variation", PLoS ONE, vol. 6, no. 12, pp. 1-20, 2011. Available: 10.1371/journal.pone.0028766 [Accessed 16 December 2019].

[2] D. Horner, W. Pirovano and G. Pesole, "Correlated substitution analysis and the prediction of amino acid structural contacts", Briefings in Bioinformatics, vol. 9, no. 1, pp. 46-56, 2007. Available: 10.1093/bib/bbm052 [Accessed 16 December 2019].

[3] T. Hopf et al., "Sequence co-evolution gives 3D contacts and structures of protein complexes", eLife, vol. 3, pp. 1-45, 2014. Available: 10.7554/elife.03430 [Accessed 16 December 2019].