Survey on Direct Coupling Analysis Method

Linh Dang

15.02.15

The main goal of this report is to overview the method named *Direct Coupling Analysis* to predict the amino acids pair contacts within single protein chain. Briefly speaking, the method of *Direct Coupling Analysis* (*DCA*) employs the *Ising* model in statistical physics to construct the interactions among site pairs and use some approximation methods to find out the most probable one. This report will be organized as following. First, we overview the background of *Ising* model. Second, we present the analogy of *Ising* model to the interaction system among amino acids pairs. After that, we summarize several approaches to estimate the parameters in *Ising* models. Finally, a modified version of *DCA* based method will be presented as to figure out the contacts between two proteins in a complex.

1 Ising Model

All the technological terms as well as definitions are based on [1] Let Λ be finite set, and denote $\Sigma_\Lambda \overset{def}{=} \{1, 2, ..., q\}^\Lambda$. In particular, q is set to 2 in *Ising* model. Suppose $A = (a_{ij})_{i,j\in\Lambda}$ is real symmetric matrix, and $\mathbf{h} = (h_i)_{i\in\Lambda}$ is a real vector. The *Hamiltonian* of the model corresponding to those parameters is a mapping $H_{A,\mathbf{h}} \colon \Sigma_\Lambda \to \mathbf{R}$ defined by

$$-H_{A,\mathbf{h}}(\sigma) \overset{def}{=} \sum_{i,j\in\Lambda} a_{ij}\sigma_i\sigma_j + \sum_{i\in\Lambda} h_i\sigma_i \tag{1}$$

and the *Gibbs* measure $G_{A,\mathbf{h}}$ on Σ_Λ is defined by

$$G_{\Lambda,A,\mathbf{h}} \overset{def}{=} \frac{1}{Z_{\Lambda,A,\mathbf{h}}} exp\left[-H_{A,\mathbf{h}}(\sigma)\right] \tag{2}$$

where $Z_{\Lambda,A,\mathbf{h}}$ is partition function, or the normalizing factor in order to make *Gibbs* measure a proper probability distribution. The partition function is defined by

$$Z_{\Lambda,A,\mathbf{h}} \overset{def}{=} \sum_{\sigma} exp\left[-H_{A,\mathbf{h}}(\sigma)\right] \tag{3}$$

2 Ising Model in Bioinformatics points of view

Important sites in protein chain defining its 3D structure tend to reserve or co-evolve over evolutionary process, which is confirmed by several researches [3–5]. The main assumption is that if two amino acids are proximity in 3D structure, they are both reserved or compensatory mutated together. The methods for detection of such co-evolved pairs have been developed for a long time and have rich literature. They could, however, be separated into two main categories which are local and global statistical probability models [2]. In local statistical methods such as *Mutual Information* inherently unable to distinguish between direct and indirect contact pairs because they consider each pair is independent to the others. In another hand, the global probability based models which take into account all pairs in calculation could overcome the shortcoming of the local based methods and give the outperformed results.

The Ising model in statistical physics could be interpreted in bioinformatics point of view as following. Let Λ be a set of all amino acids in protein chain of length $\mathcal{L}$; and like in *Ising* model, we denote $\Sigma_\Lambda \overset{def}{=} \{1, 2, ..., q\}^{|\Lambda|}$. In addition, q is set to 21 as 20 amino acid types plus gap. We notice that Σ_Λ is a set of all possible sequence of amino acids length $\mathcal{L}$. Further, consider the pairwise contacts among amino acids are modeled through a non-negative square symmetric matrix $\mathbf{J} = (\mathbf{J}_{ij})_{i,j\in\Lambda}$. And the internal interaction within single amino acid is described by a non-negative vector $\mathbf{h} = (h_i)_{i\in\Lambda}$. In the scope of structure or pairwise prediction, we are only interested in $\mathbf{J}$ which fully establish the contact information among amino acid pairs. In particular, *Direct Coupling Analysis* (*DCA*) based methods assume direct evolutionary information between two sites captured in $\mathbf{J}$. Consequently, the parameters of the model is $\mathbf{J}$ as well as $\mathbf{h}$, and the questions are how to calculate or estimate them. In order to answer this question, we look at the origin *Ising* model. The *Hamiltonian* of this protein chain is defined by

$$-H_{\mathbf{J},\mathbf{h}}(\sigma) \overset{def}{=} \sum_{i<j,i,j\in\Lambda} \mathbf{J}_{ij}\sigma_i\sigma_j + \sum_{i\in\Lambda} h_i\sigma_i \tag{4}$$

and the *Gibbs* measure on Σ_Λ is

$$\mathbf{G}_{\Lambda,\mathbf{J},\mathbf{h}} \overset{def}{=} \frac{1}{Z_{\Lambda,\mathbf{J},\mathbf{h}}} exp\left[-H_{\mathbf{J},\mathbf{h}}(\sigma)\right] \tag{5}$$

where $Z_{\Lambda,\mathbf{J},\mathbf{h}}$ as usual the partition function. We could interpret

References

[1] Erwin Bolthausen. *Spin Classes.* Springer, Berlin, germany, 2007.

[2] Marks DS, Hopf TA, and Sander C. Protein structure prediction from sequence variation. *Nat Biotech*, 30, Nov 2012.

[3] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer potts models. *Phys. Rev. E*, 87:012707, Jan 2013.

[4] Thomas A Hopf, Charlotta P I Schärfe, João P G L M Rodrigues, Anna G Green, Oliver Kohlbacher, Chris Sander, Alexandre M J J Bonvin, and Debora S Marks. Sequence co-evolution gives 3d contacts and structures of protein complexes. *eLife*, 3, 2014.

[5] Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PLoS ONE*, 6(12):e28766, 12 2011.