

Identification of novel protein domains

Abstract

A domain is a discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function. Identifying domains in a protein is instrumental in the identification of homologous proteins, thus classifying proteins into families.

When discussing proteins of the same phylogenetic background, the definition of homology is inadequate. One must define more precise terms: orthology, paralogy, and the paralogy subtypes in-paralogy and out-paralogy.¹

The rapid growth of protein sequence databases and the fact that structure and function are very hard to infer from sequence, demand a computational method for the identification of domains and classification of proteins.

There is a need to develop ways to represent domains, model a profile that defines them. One example for such a model is PSSM (position specific score matrix) which scores amino acids according to their position in the domain sequence

Another model, which is generally preferred to the latter, is HMM (hidden markov model). The model describes a state machine where any position in the sequence can transition into the next position, an insertion, or a deletion. Each such state can emit a signal – an amino acid – in a certain probability, which ultimately creates a sequence. We never know the real states, thus they are *hidden*. There are algorithms for scoring a sequence according to a model, inferring the best sequence of states that created it, and also to train a model to identify a specific motif.²

There are many database resources on the Internet providing protein family identification and annotation. The databases differ in their construction, e.g. the proteins selected as the seed multiple alignment that generates the profile, quality control, source of the profiles, and also in different features they support, such as documentation, visualization and cross links. Some examples of such databases are Pfam³ (HMM based), PROSITE (PSSM/simple pattern based), SMART (HMM based), and InterPro (combination of databases).

Using the above tools, one might create a protocol (a sequence of methods) for the identification of novel domain families, often with a prediction of their function, based on already known families.⁴

¹ Sonnhammer EL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 2002 18(12):619-620.

² Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis, Cambridge University Press, Cambridge, 1998

³ Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam Protein Families Database. *Nucleic Acids Res.* 2000 28: 263-266

⁴ Doerks T, Copley RR, Schultz J, Ponting CP, Bork P. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.* 2002 12:47-56.