Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Gleaning structural and functional information from correlations in protein multiple sequence alignments

Andrew F Neuwald

The availability of vast amounts of protein sequence data facilitates detection of subtle statistical correlations due to imposed structural and functional constraints. Recent breakthroughs using Direct Coupling Analysis (DCA) and related approaches have tapped into correlations believed to be due to compensatory mutations. This has yielded some remarkable results, including substantially improved prediction of protein intra- and inter-domain 3D contacts, of membrane and globular protein structures, of substrate binding sites, and of protein conformational heterogeneity. A complementary approach is Bayesian Partitioning with Pattern Selection (BPPS), which partitions related proteins into hierarchically-arranged subgroups based on correlated residue patterns. These correlated patterns are presumably due to structural and functional constraints associated with evolutionary divergence rather than to compensatory mutations. Hence joint application of DCA- and BPPS-based approaches should help sort out the structural and functional constraints contributing to sequence correlations.

Address
Institute for Genome Sciences and Department of Biochemistry & Molecular Biology, University of Maryland School of Medicine, 801 West Baltimore St., BioPark II, Room 617, Baltimore, MD 21201, United States

Corresponding author: Neuwald, Andrew F (aneuwald@som.umaryland.edu)

Current Opinion in Structural Biology 2016, **38**:1–8

This review comes from a themed issue on **Sequences and Topology**

Edited by **L Aravind** and **Elizabeth Meiering**

http://dx.doi.org/10.1016/j.sbi.2016.04.006

Introduction

Protein sequence data contain implicit information regarding underlying constraints important for biological function. One way to mine these data for structural and functional clues is to characterize conserved residues and statistical correlations within protein multiple sequence alignments (MSAs). The practice of extracting biological information from statistical correlations is quite old, dating at least to linkage analysis by early geneticists. Indeed, certain modern approaches are analogous to classical linkage analysis where, instead of looking for linkage between genes, one looks for linkage (i.e., couplings or correlations) between protein amino acid residues. This review focuses on recent approaches for identifying and interpreting such correlations.

Multiple sequence alignment methods

Although it may be advantageous to optimize a MSA concurrently with certain types of correlation analyses (as discussed below), nearly all programs for finding correlations in protein sequences require as input a predefined MSA, that is, one generated by another program. Since the quality of an analysis depends strongly on the quality of the input alignment, choosing the right MSA program is an important first step. Two popular state-of-the-art MSA programs used for correlation analyses are MAFFT [1] and Clustal-Ω [2]. To characterize a single protein domain, however, it is often more advantageous to start with a manually curated protein domain alignment, such as are available from the Pfam [3] database or the NCBI conserved domain database (CDD) [4]. Starting with such a MSA, or a profile hidden Markov model (HMM) derived from it, the number of aligned sequences may be expanded using the iterative search program Jackhammer [5], a web version of which is also available [6]. HHblits [7], an iterative HMM-to-HMM alignment search procedure, is also useful; in other contexts, such procedures have been found to be superior to sequence-to-profile methods for protein sequence alignment [8]. The MAPGAPS [9] program can create an alignment starting with a hierarchy of MSAs (such as are curated for the CDD), where each MSA corresponds to a subgroup within a given protein class and where the correspondence between these MSAs is defined by an alignment 'template'. MAPGAPS performs a search by creating profiles from each MSA, aligning each database sequence to its highest scoring profile, when statistically significant, and then globally aligning, as defined by the template, the conserved regions shared by all the detected sequences.

Statistical coupling analysis

The recent research described in this review was inspired, in part, by earlier work that used a weighted local mutual information approach to identify 'evolutionarily conserved pathways of energetic connectivity'—that is, sets of interacting residues mediating efficient energy conduction through a protein fold [10]. This approach, termed Statistical Coupling Analysis (SCA), starts with a covariance matrix, as do the methods discussed in the next section, and applies Principal Component Analysis (PCA) to identify groups of coevolving residue positions,

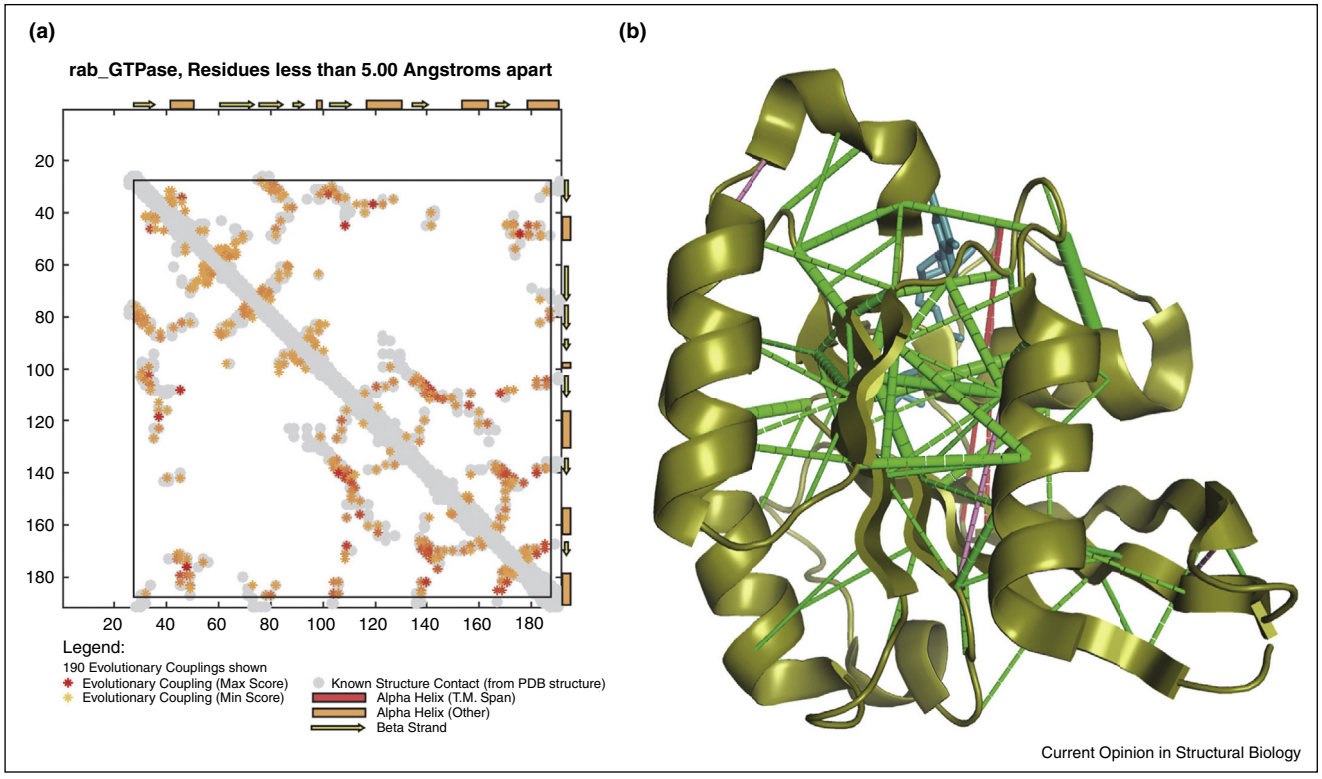
termed 'coevolving protein sectors' [11]. SCA has been used to design proteins [12] and to predict surface sites [13] and hydrophobic cavities [14] involved in allosteric regulation. A recent study [15] found that—for identification of a single sector, which includes most published SCA studies—sequence conservation alone may be used to make statistically equivalent predictions. If so, then SCA may be most useful for identifying correlations in protein alignments when multiple sectors are present. A similar approach based on multiple correspondence analysis, which is conceptually related to PCA and which was implemented in the S3det program, is designed to identify co-conserved residues responsible for subfamily-specific functions [16]. This approach defines the subfamily structure and corresponding residues simultaneously.

Inferring structural interactions from correlated residues

Identifying structural constraints from residue–residue correlations has been a topic of study for some time (e.g., see references in [17,18••]) and involves analysis of a covariance matrix derived from how often the various pairs of amino acid residues occur at each pair of positions in a MSA. The rationale for this is that mutations occurring at one residue position often result in compensatory mutations at other, structurally interacting residue positions. A problem with this straightforward approach, however, is that residue positions may be correlated transitively; that is, if residue position i interacts with position j and j with position k, then residues at positions i and k may be correlated even though they fail to structurally interact directly. A critical breakthrough in this area came with the development of two methods, Direct Coupling Analysis (DCA) [19••] (Figure 1) and sparse inverse covariance estimation [20], which distinguish direct from indirect correlations by inverting the covariance matrix (for in depth reviews see [21,22,23•]). A further improvement in the DCA approach involved using pseudo-likelihood maximization [24] to calculate the coupling parameters rather than the original mean field approximation. Other improvements have also been reported based on multivariate Gaussian modeling [25] and on a 3-step procedure [26]. Downloadable programs implementing these approaches include PconsFold [27], PSICOV [20], CCMpred [28], MetaPSICOV [29] and FreeContact [30].

Some remarkable results have been achieved using these approaches. Recently, for example, structural and functional insights have been gained into membrane proteins

Figure 1



Direct Coupling Analysis (DCA) of Rab11a GTPase. This output was obtained from the web-based EVcouplings program (http://EVfold.org). **(a)** Map of the highest scoring coupled residue pairs compared to the native contacts. **(b)** The top predicted contacts shown as green lines and out of range predicted contacts as red lines within a Rab11a structure (pdb_id: 1oiw) [71].