Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

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

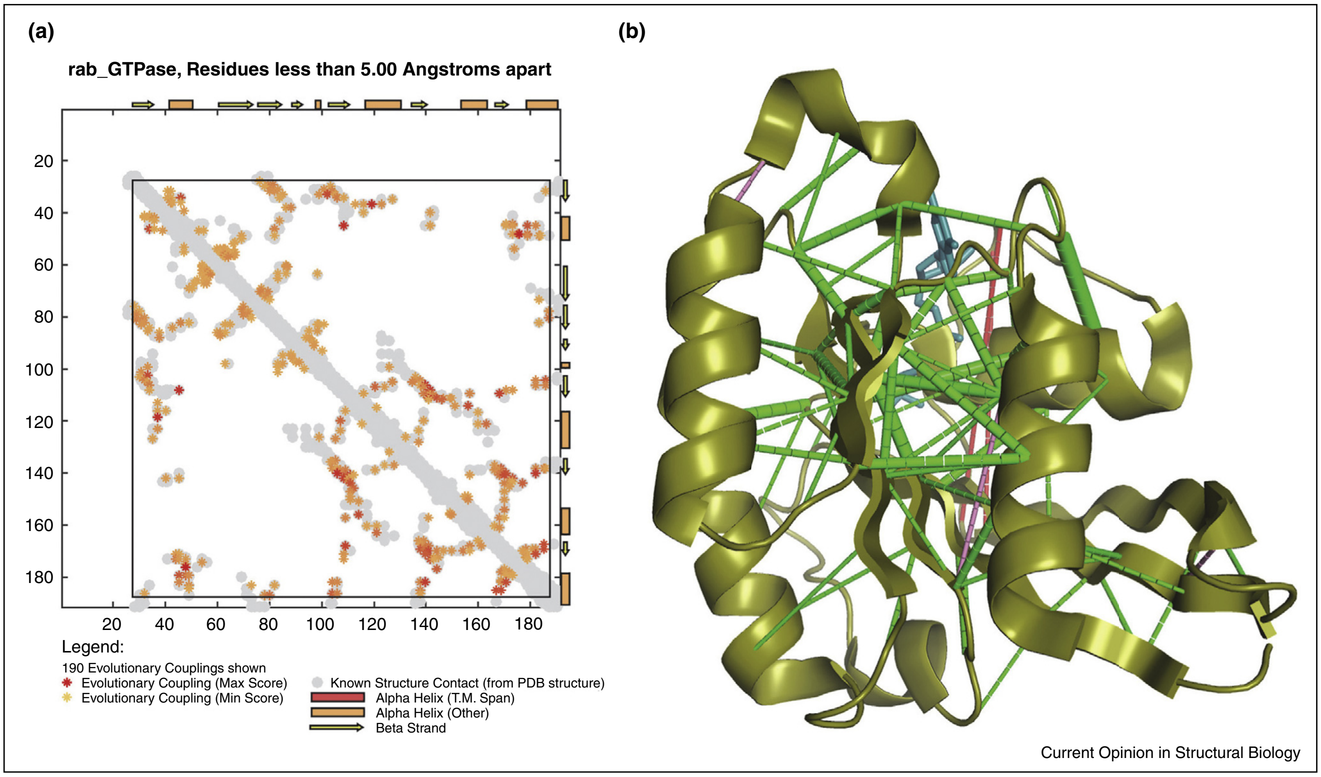
The recent research described in this review was inspired, in part, by earlier work that used a weighted local mutual information approach to identify 'evolutionarily conserved pathways of energetic connectivity'—that is, sets of interacting residues mediating efficient energy conduction through a protein 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



Current Opinion in Structural Biology

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].

[31••], the structures of which are difficult to determine through crystallography. In particular, this has led to structural insights regarding odor binding and ion conduction domains within insect odorant receptors [32], to 3D-structure predictions for 19 transmembrane β-barrel proteins [33], and to predictions of functionally relevant residues in the *E. coli* β-barrel protein BamA [34]. Direct coupling analysis also has been used to determine residue interactions between internal repeats [35] and within protein complexes [36,37]. An antiparallel homophilic interface seen in a crystal structure of extracellular cadherin domains was supported by evolutionary covariance analysis [38]. DCA was used to study how bacterial two component systems maintain their ability to transmit signals with high specificity and potentially how to rationally redesign these systems [39]. Hence DCA-related approaches have been validated through numerous studies.

Other studies have compared these newest, direct coupling methods with more traditional mutual information (MI) approaches. A comparison of direct coupling versus MI methods for detection of inter-protein contacts [40] confirmed the former's generally superior performance. Mao *et al.* [41] likewise confirmed the superiority of DCA for detecting tertiary structural contacts. However, Clark *et al.* [42] report that two multidimensional extensions of MI methods (mdMI), which are designed to remove the effect of ternary/quarternary interdependencies, are comparable to the newest direct coupling pseudolikelihood methods—though these approaches shared less than 65% overlap between their top scoring residue pairs. Another study casts doubt on the assumption that observed patterns of covariation are caused by molecular coevolution; that is, whether mutations at one site impose evolutionary pressures at neighboring sites [43•]. The authors argue that, because most methods are tree-independent, their results are difficult to interpret evolutionarily. They report that covariation may be due to rare independent changes at conserved sites as well as to correlated changes resulting from coevolution.

Hybrid approaches to structure prediction

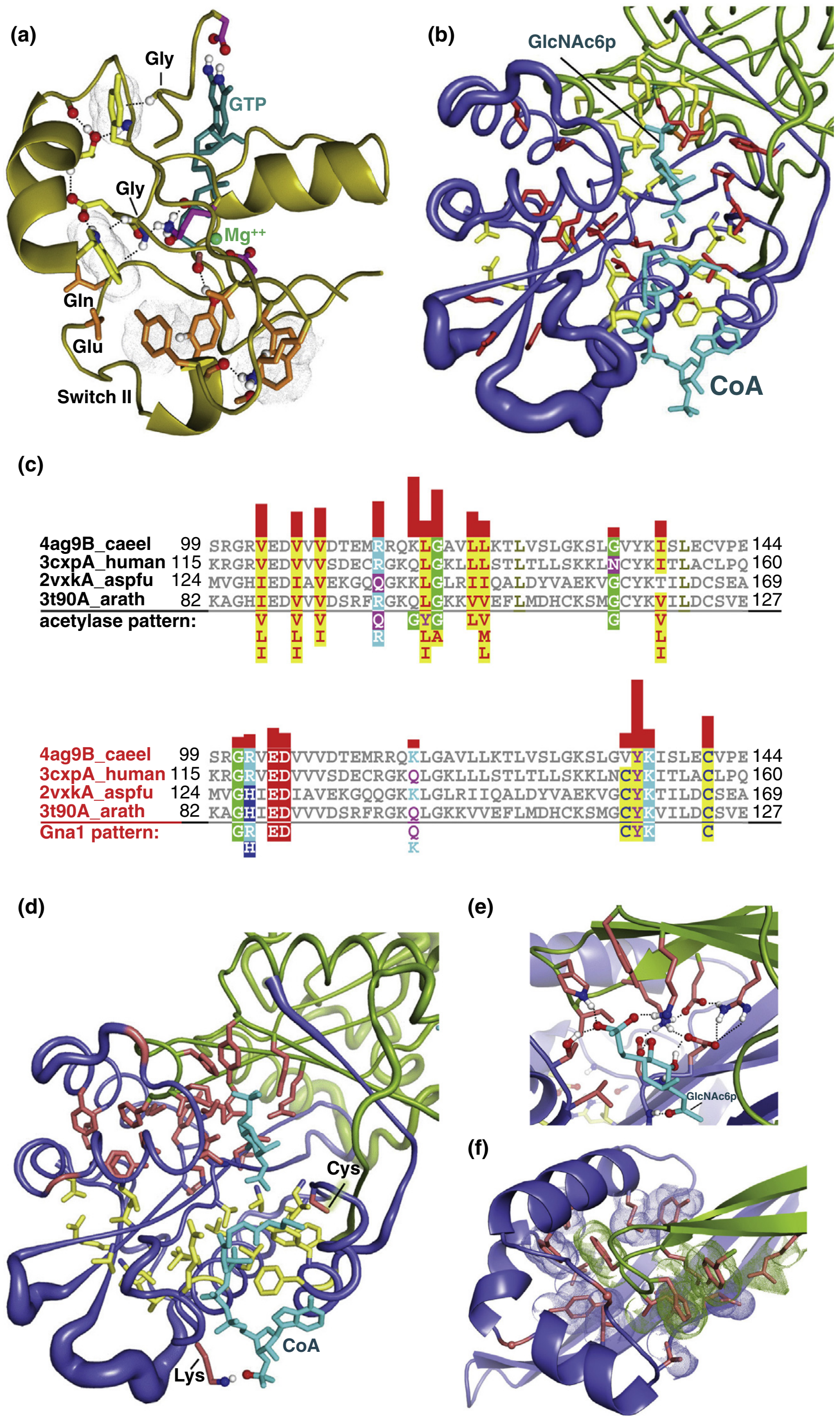
Evolutionary covariance analysis has been integrated into other protein structure analysis methods. It has been combined with molecular dynamics (MD) simulations to explore protein conformational heterogeneity by converting contact predictions into an ensemble of structural states for a given protein [44]. A similar combined DCA-MD approach was used to explore how the response regulator of two-component signal transduction systems transmits the activation signal between its N-terminal receiver and C-terminal effector domains [45]. However, when DCA was applied to Hsp70 chaperones without MD simulations, it still captured the large-scale conformational transitions characteristic of these proteins and predicted a functional homodimeric state [46•]. The Baker group has integrated their co-evolution-based contact prediction program, GREMLIN [47], into their Rosetta structure prediction program [48] to model the Zinc transporter hZIP4 [49]. They applied this hybrid approach to large scale determination of unsolved protein structures [50•], which they have made publically available; for two proteins this approach has resulted in unprecedented accuracy in *de novo* structure prediction for the CASP11 blind test [51]. The Jones group has similarly integrated their PSICOV program into their FRAGFOLD program to improve de novo structure prediction [52]. The GREMLIN program has also been combined with a physicochemical approach to structure prediction [53] that is available over a web server [54]. Evolutionary coupling has improved structure determination by NMR spectroscopy [55•]. DCA seems likely to become an integral component of de novo structure prediction.

Bayesian partitioning with pattern selection

An alternative approach for inferring biological information from MSA correlations and a focus of this review is Bayesian Partitioning with Pattern Selection (BPPS) [56,57]. Rather than focusing on a covariance or mutual information matrix based on residue pairs, BPPS focuses on correlations more loosely defined as dependencies involving many residue positions. It uses Markov chain Monte Carlo (MCMC) sampling to partition a MSA into subgroups, each of which is defined by a correlated residue pattern that best distinguishes the sequences in that subgroup from other, closely-related sequences. Hence, each pattern consists of an arbitrary number of correlated residue positions. An underlying assumption is that pattern residues encode protein properties shared by members of the corresponding subgroup. A recent 'multiple-category BPPS' sampler [58•,59] automatically arranges a MSA hierarchically into subgroups by searching for the mode of the posterior probability distribution over all such hierarchies. For a major protein class the MSA typically contains at least 100,000 sequences, which are far too many for covariance matrix-based methods that identify function-related conserved patterns, such as S3det cited above. After BPPS sampling partitions a protein class into functionally divergent subgroups, an auxiliary program maps correlated residue patterns to available protein structures as an aid to biological interpretation. Such analyses are currently non-trivial to perform. Hence, to make this approach widely available, precomputed BPPS analyses are being incorporated into the NCBI CDD [60].

Figure 2a illustrates a key aspect of BPPS analysis of P-loop GTPases. It focuses on the structural locations of the most discriminating pattern residues identified for three hierarchically-arranged subgroups to which Rab11A GTPase belongs, namely all P-loop GTPases, Ras-like GTPases, and a subgroup consisting of Rab, Rho and Ran GTPases [57]. The P-loop GTPase conserved residues

Figure 2


(a)
Gly
GTP
Gly
Mg++
Gln
Glu
Switch II
(b)
GlcNAc6p
CoA
(c)
4ag9B_caeel 99 SRGRVEDVVVDTEMRRQKLGAVLLKTLVSLGKSLGVYKISLECVPE 144
3cxpA_human 115 KRGRVEDVVVSDECRGKQLGKLLLSTLTLLSKKLNCYKITLACLPQ 160
2vxkA_aspfu 124 MVGHIEDIAVEKGQQGKKLGLRIIQALDYVAEKVGCYKTILDCSEA 169
3t90A_arath 82 KAGHIEDVVVDSRFRGKQLGKKVVEFLMDHCKSMGCYKVILDCSVE 127
acetylase pattern:
V V V Q GYG LV G V
L L I R LA M L
I I I L I
4ag9B_caeel 99 SRGRVEDVVVDTEMRRQKLGAVLLKTLVSLGKSLGVYKISLECVPE 144
3cxpA_human 115 KRGRVEDVVVSDECRGKQLGKLLLSTLTLLSKKLNCYKITLACLPQ 160
2vxkA_aspfu 124 MVGHIEDIAVEKGQQGKKLGLRIIQALDYVAEKVGCYKTILDCSEA 169
3t90A_arath 82 KAGHIEDVVVDSRFRGKQLGKKVVEFLMDHCKSMGCYKVILDCSVE 127
Gna1 pattern:
GR ED Q CYK C
H K
(d)
Cys
CoA
Lys
(e)
GlcNAc6p
(f)

Current Opinion in Structural Biology

BPPS analysis of correlated residue patterns. See discussion in BPPS subsection. **(a)** Locations of top scoring pattern residues in human Rab11a GTPase complexed with GTPγS (pdb_id: 1oiw) [71]. Residue sidechains most characteristic of P-loop GTPases, Ras-like GTPases and Rho/Rab/Ran GTPases are shown in magenta, orange and yellow, respectively. **(b–f)** FRpred and BPPS analyses of Gna1 acetyltransferases interpreted in light of the homodimeric structure of *C. elegans* glucosamine-6-phosphate *N*-acetyltransferase (Gna1) complexed with coenzyme A (CoA) and *N*-acetylglucosamine-6-phosphate (GlcNA6p) (pdb_id: 4ag9) [68]. **(b)** Structural locations of Gna1 residues identified and classified by FRpred [67]. Color scheme: the backbones of the two subunits within the homodimer, blue and green; conserved residue sidechains, yellow; subtype residues, red; mixed residues, orange; CoA and substrate, cyan. FRpred fails to identify both several residues interacting with substrate, as shown in **(e)**, are known to be involved in binding GTP or GDP. Ras-like GTPases function as on-off switches in signaling pathways; they are turned on when bound to GTP and turned off when bound to GDP. Two of the BPPS-identified Ras-like pattern residues, which are labeled as Gln and Glu in Figure 2a, are proposed (based on experimental studies) to be involved in hydrolysis of GTP to GDP and in exchange of GTP for GDP, respectively. These are located in the Switch II region, which undergoes conformational changes associated with signal transduction. Residues at five other positions in the Ras-like pattern, first identified through BPPS analysis [61], mutually-interact near the C-terminal end of the switch II region. Within available crystal structures these residues, which form two distinct conformations, were implicated in the on/off switching mechanism [61]. The four pattern residues most distinctive of Rab/Rho/Ran GTPases form aromatic CH–π interactions proposed to stabilize two glycine residue 'flexible hinges' within guanine nucleotide binding loops; the pattern residues were hypothesized to function as a 'glycine brace' facilitating nucleotide binding and/or release [62]. These findings are quite distinct from those obtained through an SCA analysis of GTPases [63], illustrating how the BPPS and SCA approaches address distinct problems. BPPS and other correlated residue analyses likewise differ from phylogenetic-tree-based functional residue prediction methods, which are not covered in this review.

As noted recently [64,65], the benchmarking of programs that identify sequence determinants of protein function is problematic because experimental studies defining the biochemical functions of specific residues are incomplete. Consequently, identified residues involved in important but uncharacterized functions will be scored incorrectly as false positives. For this reason, BPPS analysis focuses on identifying the statistically most striking correlated residue patterns, which it uses to define functionally divergent subgroups. Success is evaluated by identifying pattern-defined subgroups in simulated data (where the true solution is known) and on assessing the robustness, reproducibility and stochastic uncertainty of results for real data [66]. In principle, the statistical significance of 'surprising' results might be assessed by computing a P-value. However, how to adjust for multiple hypotheses by determining the number of equally surprising results is unclear. Nevertheless, in the light of other information, biological significance may be assessed qualitatively based on 'interpretability'. This is illustrated through the following BPPS analysis of GNAT acetyltransferases with a focus on glucosamine-6-phosphate *N*-acetyltransferase (Gna1). In contrast to residues identified using another (web-based) functional residue prediction program [67] (Figure 2b), BPPS identifies two categories of pattern residues (Figure 2c) that are structurally partitioned (in a strikingly non-random manner) into two subdomains (Figure 2d). One subdomain, which harbors residues generally shared by all acetyltransferases, is involved in binding to coenzyme A (CoA). The other subdomain, which harbors residues characteristic of the Gna1 subgroup, is involved both in substrate binding (Figure 2e) and in the formation of a homodimeric interface adjacent to the substrate-binding site (Figure 2f). In addition, two Gna1 pattern residues occur within the CoA-binding subdomain (Figure 2d): (i) a cysteine residue that is located near the active site and that forms a disulfide bond with CoA [68], thereby suggesting a functional role, and (ii) a lysine residue that hydrogen bonds to a CoA phosphate group and thus may facilitate catalysis by positioning CoA. Hence, in this case the BPPS analysis is readily interpretable in the light of structural and biochemical information.

Toward comprehensive modeling of a protein class

Ongoing research is extending BPPS analysis to fit a hierarchical model to those sequences observed for an entire major protein domain class and to thereby define a likelihood distribution over sequence space for that class. This requires, of course, that the observed sequences be sufficiently abundant and representative of the class. This distribution is typically quite complex—consisting of a main probability density cloud (corresponding to the entire class) within which are multiple subclouds (superfamilies), and, within these, sub–subclouds (families), *etc.* Bayesian MCMC sampling is widely recognized as the most effective approach for characterizing such a complex, high dimensional distribution. For this, BPPS and MSA statistical models are being combined into a single, coherent BPPS/MSA model that captures both residue frequencies at each position and residue correlations. A MSA sampler that is statistically consistent with the BPPS sampler has been recently developed [69]. To avoid modeling extraneous features and random noise the Minimum Description Length (MDL) principle [70] is applied to adjust for the number of models implicitly

(Figure 2 Legend Continued) and the cysteine residue indicated in **(d)**. The biological interpretation of these results is less clear than that of the following BPPS analysis. **(c)** BPPS analysis of 200,028 acetylase domains with a focus on the Gna1 subgroup (979 proteins). The top and bottom alignments highlight, respectively, the residues generally conserved in all acetyltransferases and the correlated residues defining the Gna1 subgroup. Aligned sequences correspond to four Gna1 acetyltransferases of known structure; for clarity, only a subregion of the alignment is shown. The heights of the red bars above the alignments correspond to the BPPS scores at pattern positions. **(d)** Structural locations of BPPS pattern residues within Gna1. Residues sidechains most characteristic of *N*-acetyltransferases and of the Gna1 subgroup are colored yellow and red, respectively. A cysteine residue that forms a disulfide bond with CoA [68] and a lysine residue that forms a hydrogen bond with a phosphate group of CoA are indicated. **(e)** Close up of BPPS pattern residues involved in GlcNA6p binding. **(f)** Close up of pattern residues forming a homodimeric interface near the substrate binding pocket. Figures were created using PyMOL, Schrödinger, LLC.

considered during sampling. This ensures that sequence regions are aligned and that correlated residue patterns and corresponding subgroups are defined only when justified statistically.

Together, these (and others) features of BPPS/MSA sampling allow it, simultaneously, to (i) define a hierarchy of divergent subgroups based on correlated residue patterns, (ii) accurately align each subgroup, and (iii) define a posterior probability distribution for the protein class. From this, lineage-specific sequence and structural information may be extracted for each subgroup, as illustrated for Gna1 in Fig. 2c–f. By comparing and contrasting such lineage-specific perspectives for different subgroups, a broad functional and structural understanding of an entire protein class can be obtained. Finally, because residue correlations are due to both compensatory mutations and functional divergence, prior BPPS/MSA partitioning into divergent subgroups may aid DCA-based prediction of 3D structural contacts within each subgroup. Thus combining BPPS analysis with DCA, SCA and other correlated residue analyses is an area of active investigation.

Conflicts of interest statement

There are no conflicts of interest.

Acknowledgements

I thank Stephen F. Altschul for critical reading of the manuscript and for helpful suggestions. This work was supported by funding from the University of Maryland, Baltimore.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Katoh K, Standley DM: **MAFFT: iterative refinement and additional methods**. *Methods Mol Biol* 2014, **1079**:131-146.

2. Sievers F, Higgins DG: **Clustal omega, accurate alignment of very large numbers of sequences**. *Methods Mol Biol* 2014, **1079**:105-116.

3. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL *et al.*: **Pfam: the protein families database**. *Nucleic Acids Res* 2014, **42**:D222-D230.

4. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ *et al.*: **CDD: NCBI's conserved domain database**. *Nucleic Acids Res* 2015, **43**:D222-D226.

5. Eddy SR: **A new generation of homology search tools based on probabilistic inference**. *Genome Inf* 2009, **23**:205-211.

6. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR: **HMMER web server: 2015 update**. *Nucleic Acids Res* 2015, **43**:W30-W38.

7. Remmert M, Biegert A, Hauser A, Soding J: **HHblits: lightning-fast iterative protein sequence searching by HMM–HMM alignment**. *Nat Methods* 2012, **9**:173-175.

8. Yan R, Xu D, Yang J, Walker S, Zhang Y: **A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction**. *Sci Rep* 2013, **3**:2619.

9. Neuwald AF: **Rapid detection, classification and accurate alignment of up to a million or more related protein sequences**. *Bioinformatics* 2009, **25**:1869-1875.

10. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families**. *Science* 1999, **286**:295-299.

11. Halabi N, Rivoire O, Leibler S, Ranganathan R: **Protein sectors: evolutionary units of three-dimensional structure**. *Cell* 2009, **138**:774-786.

12. Reynolds KA, Russ WP, Socolich M, Ranganathan R: **Evolution-based design of proteins**. *Methods Enzymol* 2013, **523**:213-235.

13. Reynolds KA, McLaughlin RN, Ranganathan R: **Hot spots for allosteric regulation on protein surfaces**. *Cell* 2011, **147**:1564-1575.

14. Tanwar AS, Goyal VD, Choudhary D, Panjikar S, Anand R: **Importance of hydrophobic cavities in allosteric regulation of formylglycinamide synthetase: insight from xenon trapping and statistical coupling analysis**. *PLOS ONE* 2013, **8**:e77781.

15. Tesileanu T, Colwell LJ, Leibler S: **Protein sectors: statistical coupling analysis versus conservation**. *PLoS Comput Biol* 2015, **11**:e1004091.

16. Rausell A, Juan D, Pazos F, Valencia A: **Protein interactions and ligand binding: from protein subfamilies to functional specificity**. *Proc Natl Acad Sci U S A* 2010, **107**:1995-2000.

17. de Juan D, Pazos F, Valencia A: **Emerging methods in protein co-evolution**. *Nat Rev Genet* 2013, **14**:249-261.

18. •• Cocco S, Monasson R, Weigt M: **From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction**. *PLoS Comput Biol* 2013, **9**:e1003176.

This well-written paper introduces a Hopfield–Potts model for interpolating between principle component analysis, which identifies the most correlated residues, and direct coupling analysis, which aims at predicting residue–residue contacts based on the maximum entropy principle. This is an excellent read for better understanding both the distinctions between methods and the mathematics underlying covariance analysis of multiple-sequence alignments.

19. •• Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M: **Direct-coupling analysis of residue coevolution captures native contacts across many protein families**. *Proc Natl Acad Sci U S A* 2011, **108**:E1293-E1301.

This is a landmark paper on direct coupling analysis.

20. Jones DT, Buchan DW, Cozzetto D, Pontil M: **PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments**. *Bioinformatics* 2012, **28**:184-190.

21. Marks DS, Hopf TA, Sander C: **Protein structure prediction from sequence variation**. *Nat Biotechnol* 2012, **30**:1072-1080.

22. Morcos F, Hwa T, Onuchic JN, Weigt M: **Direct coupling analysis for protein contact prediction**. *Methods Mol Biol* 2014, **1137**:55-70.

23. • Stein RR, Marks DS, Sander C: **Inferring pairwise interactions from biological data using maximum-entropy probability models**. *PLoS Comput Biol* 2015, **11**:e1004182.

This review describes the mathematics underlying maximum entropy-based methods for inferring direct interactions from biological data. The authors show how these methods can be more generally applied to biological problems beyond protein 3D structure prediction.

24. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ: **Learning generative models for protein fold families**. *Proteins* 2011, **79**:1061-1078.

25. Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A: **Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners**. *PLOS ONE* 2014, **9**:e92721.

26. Feinauer C, Skwark MJ, Pagnani A, Aurell E: **Improving contact prediction along three dimensions**. *PLoS Comput Biol* 2014, **10**:e1003847.

27. Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A: **PconsFold: improved contact predictions improve protein models**. *Bioinformatics* 2014, **30**:i482-i488.

28. Seemayer S, Gruber M, Soding J: **CCMpred – fast and precise prediction of protein residue–residue contacts from correlated mutations**. *Bioinformatics* 2014, **30**:3128-3130.

29. Jones DT, Singh T, Kosciolek T, Tetchner S: **MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins**. *Bioinformatics* 2015, **31**:999-1006.

30. Kajan L, Hopf TA, Kalas M, Marks DS, Rost B: **FreeContact: fast and free software for protein contact prediction from residue co-evolution**. *BMC Bioinf* 2014, **15**:85.

31. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS:
•• **Three-dimensional structures of membrane proteins from genomic sequencing**. *Cell* 2012, **149**:1607-1621.

The authors use residue covariation to predict previously unknown 3D structures for 11 transmembrane proteins from sequence alone. The unprecedented accuracy of such predictions was confirmed through *de novo* computation of transmembrane proteins of known structure from 23 families.

32. Hopf TA, Morinaga S, Ihara S, Touhara K, Marks DS, Benton R: **Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors**. *Nat Commun* 2015, **6**:6077.

33. Hayat S, Sander C, Marks DS, Elofsson A: **All-atom 3D structure prediction of transmembrane beta-barrel proteins from sequences**. *Proc Natl Acad Sci U S A* 2015, **112**: 5413-5418.

34. Dwyer RS, Ricci DP, Colwell LJ, Silhavy TJ, Wingreen NS: **Predicting functionally informative mutations in *Escherichia coli* BamA using evolutionary covariance analysis**. *Genetics* 2013, **195**:443-455.

35. Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU: **Capturing coevolutionary signals in repeat proteins**. *BMC Bioinf* 2015, **16**:207.

36. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS: **Sequence co-evolution gives 3D contacts and structures of protein complexes**. *Elife* 2014, **3**.

37. dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN: **Dimeric interactions and complex formation using direct coevolutionary couplings**. *Sci Rep* 2015, **5**:13652.

38. Nicoludis JM, Lau SY, Scharfe CP, Marks DS, Weihofen WA, Gaudet R: **Structure and sequence analyses of clustered protocadherins reveal antiparallel interactions that mediate homophilic specificity**. *Structure* 2015, **23**:2087-2098.

39. Cheng RR, Morcos F, Levine H, Onuchic JN: **Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information**. *Proc Natl Acad Sci U S A* 2014, **111**:E563-E571.

40. Avila-Herrera A, Pollard KS: **Coevolutionary analyses require phylogenetically deep alignments and better null models to accurately detect inter-protein contacts within and between species**. *BMC Bioinf* 2015, **16**:268.

41. Mao W, Kaya C, Dutta A, Horovitz A, Bahar I: **Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution**. *Bioinformatics* 2015, **31**: 1929-1937.

42. Clark GW, Ackerman SH, Tillier ER, Gatti DL: **Multidimensional mutual information methods for the analysis of covariation in multiple sequence alignments**. *BMC Bioinf* 2014, **15**:157.

43. Talavera D, Lovell SC, Whelan S: **Covariation is a poor measure**
• **of molecular coevolution**. *Mol Biol Evol* 2015, **32**:2456-2468.

Based on theoretical and empirical analysis the authors cast doubt on the assumption that covariation patterns are caused by compensatory mutations and co-evolution. While not questioning the significance of recent breakthroughs in 3D contact predictions, these authors challenge us to rethink the reasons for the success of these methods.

44. Sutto L, Marsili S, Valencia A, Gervasio FL: **From residue coevolution to protein conformational ensembles and functional dynamics**. *Proc Natl Acad Sci U S A* 2015, **112**: 13567-13572.

45. Ahmad A, Cai Y, Chen X, Shuai J, Han A: **Conformational dynamics of response regulator RegX3 from *Mycobacterium tuberculosis***. *PLOS ONE* 2015, **10**:e0133389.

46. Malinverni D, Marsili S, Barducci A, De Los Rios P: **Large-scale**
• **conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones**. *PLoS Comput Biol* 2015, **11**:e1004262.

This study shows how DCA contact predictions for Hsp70 chaperones, which are inconsistent with any single conformational form, are explained by the 3D contacts established in alternative conformational forms observed in available Hsp70 crystal structures.

47. Kamisetty H, Ovchinnikov S, Baker D: **Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era**. *Proc Natl Acad Sci U S A* 2013, **110**:15674-15679.

48. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D: **High-resolution comparative modeling with RosettaCM**. *Structure* 2013, **21**:1735-1742.

49. Antala S, Ovchinnikov S, Kamisetty H, Baker D, Dempski RE: **Computation and functional studies provide a model for the structure of the zinc transporter hZIP4**. *J Biol Chem* 2015, **290**:17796-17805.

50. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H,
• Grishin NV, Baker D: **Large-scale determination of previously unsolved protein structures using evolutionary information**. *Elife* 2015, **4**.

This paper describes *de novo* blind structure predictions of unprecedented accuracy for two proteins using a combination of residue–residue co-evolutionary information and the Rosetta structure prediction program. The authors applied this approach to generate structural models for 58 prokaryotic protein families lacking 3D structures, examination of which led to mechanistic and functional hypotheses.

51. Monastyrskyy B, D'Andrea D, Fidelis K, Tramontano A, Kryshtafovych A: **New encouraging developments in contact prediction: assessment of the CASP11 results**. *Proteins* 2015.

52. Kosciolek T, Jones DT: **De novo structure prediction of globular proteins aided by sequence variation-derived contacts**. *PLOS ONE* 2014, **9**:e92197.

53. Schneider M, Brock O: **Combining physicochemical and evolutionary information for protein contact prediction**. *PLOS ONE* 2014, **9**:e108438.

54. Mabrouk M, Putz I, Werner T, Schneider M, Neeb M, Bartels P, Brock O: **RBO Aleph: leveraging novel information sources for protein structure prediction**. *Nucleic Acids Res* 2015, **43**: W343-W348.

55. Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT:
• **Protein structure determination by combining sparse NMR data with evolutionary couplings**. *Nat Methods* 2015, **12**: 751-754.

This study demonstrates how residue–residue covariance information can complement NMR data for determining protein structures. The authors provide a detailed description of how to apply this approach.

56. Neuwald AF, Kannan N, Poleksic A, Hata N, Liu JS: **Ran's C-terminal, basic patch and nucleotide exchange mechanisms in light of a canonical structure for Rab, Rho, Ras and Ran GTPases**. *Genome Res* 2003, **13**:673-692.

57. Neuwald AF: **Surveying the manifold divergence of an entire protein class for statistical clues to underlying biochemical mechanisms**. *Stat Appl Genet Mol Biol* 2011, **10**:36.

58. Neuwald AF: **A Bayesian sampler for optimization of protein**
• **domain hierarchies**. *J Comput Biol* 2014, **21**:269-286.

This paper describes the statistics and algorithm underlying the BPPS sampler for hierarchically classifying protein sequences based on correlated residue patterns. This approach, which builds upon earlier work, is typically applied to very large alignments of more than 100,000 sequences.

59. Neuwald AF: **Protein domain hierarchy Gibbs sampling strategies**. *Stat Appl Genet Mol Biol* 2014, **13**:497-517.

60. Neuwald AF, Lanczycki CJ, Marchler-Bauer A: **Automated hierarchical classification of protein domain subfamilies based on functionally-divergent residue signatures**. *BMC Bioinf* 2012, **13**:144.

61. Neuwald AF: **The charge-dipole pocket: a defining feature of signaling pathway GTPase on/off switches**. *J Mol Biol* 2009, **390**:142-153.

62. Neuwald AF: **The glycine brace: a component of Rab, Rho, and Ran GTPases associated with hinge regions of guanine- and phosphate-binding loops**. *BMC Struct Biol* 2009, **9**:11.

63. Hatley ME, Lockless SW, Gibson SK, Gilman AG, Ranganathan R: **Allosteric determinants in guanine nucleotide-binding proteins**. *Proc Natl Acad Sci U S A* 2003, **100**:14445-14450.

64. Dessimoz C, Skunca N, Thomas PD: **CAFA and the open world of protein function predictions**. *Trends Genet* 2013, **29**:609-610.

65. Jiang Y, Clark WT, Friedberg I, Radivojac P: **The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective**. *Bioinformatics* 2014, **30**:i609-i616.

66. Neuwald AF: **Evaluating, comparing, and interpreting protein domain hierarchies**. *J Comput Biol* 2014, **21**:287-302.

67. Fischer JD, Mayer CE, Soding J: **Prediction of protein functional residues from sequence by probability density estimation**. *Bioinformatics* 2008, **24**:613-620.

68. Dorfmueller HC, Fang W, Rao FV, Blair DE, Attrill H, van Aalten DM: **Structural and biochemical characterization of a trapped coenzyme A adduct of *Caenorhabditis elegans* glucosamine-6-phosphate N-acetyltransferase 1**. *Acta Crystallogr D Biol Crystallogr* 2012, **68**:1019-1029.

69. Neuwald AF, Altschul SF: **Bayesian top-down protein sequence alignment with inferred position-specific gap penalties**. *PLoS Comput Biol* 2016, **12**:e1004936 http://dx.doi.org/10.1371/journal.pcbi.1004936.

70. Grunwald PD: *The Minimum Description Length Principle*. Boston: MIT Press; 2007.

71. Pasqualato S, Senic-Matuglia F, Renault L, Goud B, Salamero J, Cherfils J: **The structural GDP/GTP cycle of Rab11 reveals a novel interface involved in the dynamics of recycling endosomes**. *J Biol Chem* 2004, **279**:11480-11488.