Sequence analysis

eCOMPASS: evaluative comparison of multiple protein alignments by statistical score

Andrew F. Neuwald[1,*], Bryan D. Kolaczkowski[2] and Stephen F. Altschul[3]

[1]Department of Biochemistry & Molecular Biology, University of Maryland School of Medicine, Baltimore, MD 21201, USA, [2]Department of Microbiology & Cell Science, University of Florida, Gainesville, FL 32611, USA and [3]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Detecting subtle biologically relevant patterns in protein sequences often requires the construction of a large and accurate multiple sequence alignment (MSA). Methods for constructing MSAs are usually evaluated using benchmark alignments, which, however, typically contain very few sequences and are therefore inappropriate when dealing with large numbers of proteins.

Results: eCOMPASS addresses this problem using a statistical measure of relative alignment quality based on direct coupling analysis (DCA): To maintain protein structural integrity over evolutionary time, substitutions at one residue position typically result in compensating substitutions at other positions. eCOMPASS computes the statistical significance of the congruence between high scoring directly coupled pairs and 3D contacts in corresponding structures, which depends upon properly aligned homologous residues. We illustrate eCOMPASS using both simulated and real MSAs.

Availability and Implementation: The eCOMPASS executable, C++ open source code and input data sets are available at https://www.igs.umaryland.edu/labs/neuwald/software/compass.

Contact: aneuwald@som.umaryland.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein sequence analyses, and particularly those that are statistically based, often rely upon very large multiple sequence alignments (MSAs), consisting of tens or hundreds of thousands of sequences belonging to a large superfamily. Using such an alignment increases the statistical power and breadth of an analysis and, by partitioning the MSA into hierarchically arranged subgroups based on subgroup-specific patterns (Neuwald, 2014), one can identify sequence and structural features likely determining functional specificity. For example, this approach has been used (Neuwald, et al., 2012) to automate the manual curation of hierarchical MSAs (hiMSAs) for the NCBI Conserved Domain Database (CDD) (Yang, et al., 2020) and, when applied to an MSA of 474,040 AAA+ ATPases, has revealed sequence and structural properties implicated in DNA clamp loader functional specificity (Tondnevis, et al., 2020). We have performed similar analyses using alignments of 237,359 N-acetyltransferases, 127,418 GTPases, 131,321 helicases, 45,799 exonuclease-endonuclease-phosphatases and 23,592 DNA glycosylases (Neuwald, et al., 2018) and of 33,760 TIR domains (Toshchakov and Neuwald, 2020). It is important, of course, that such alignments be as biologically accurate as possible. However, it is well known that only heuristic methods are available for constructing even small alignments, and these produce results that may be far from optimal (Edgar, 2010). Generally, an MSA method's accuracy is evaluated using a set of benchmark alignments that are manually curated using structural data and each of which typically contain relatively few sequences. However, there are many potential problems with these evaluations. First, they rely upon the accuracy of the benchmark alignments, which may itself be in question (Ashkenazy, et al., 2019; Fletcher and Yang, 2010; Kim and Lee, 2007; Levy Karin, et al., 2014; Thompson, et al., 2011). Second, they implicitly assume the accuracy of an MSA on a benchmark set of sequences is a good

proxy for its accuracy on a much larger superset. This may not be the case, particularly when the larger set contains many protein subgroups within a superfamily, not all of which are represented within the benchmark alignment. Curating large benchmark MSAs is error prone and may be prohibitively labor intensive. Finally, the relative accuracy of one MSA method to another on a set of benchmark alignments is no guarantee that it will produce the more accurate alignment for a specific set of sequences of interest, particularly one that is large and diverse.

We define an accurate alignment to be one that reflects sequence homology. A more accurate MSA should reveal evolutionarily conserved structural and functional constraints better than a less accurate one. In large, diverse sequence sets such constraints become more statistically evident, thereby allowing subtly conserved homologous regions to be identified and aligned, as illustrated in (Neuwald and Hirano, 2000; Neuwald and Poleksic, 2000).

Because obtaining a highly accurate MSA typically requires manual curation, we have developed and applied the Multiply Aligned Profiles for Global Alignment of Protein Sequences (MAPGAPS) program (Neuwald, 2009), which uses a manually curated hiMSA as a query to identify and align database sequences belonging to a modeled superfamily. Within a hiMSA each subgroup alignment is profiled and aligned to the other subgroup alignments. Using this feature, MAPGAPS creates an MSA with accuracy comparable to that of the hiMSA (Neuwald, et al., 2020). This assumes that each subgroup is accurately aligned both internally and relative to other subgroups, which is typically not yet the case. Hence, to further improve this approach, we need to assess alignment quality for each subgroup and for the MSA as a whole.

Here we introduce eCOMPASS, a program that evaluates the relative accuracy of two MSAs of the same large set of sequences by applying direct coupling analysis (DCA) based upon pseudo-likelihood maximization in conjunction with a procedure to estimate statistical significance. It requires as input only the MSAs themselves and structural coordinates for a minimum number (ideally at least ten) of the aligned sequences. It does not rely upon any set of benchmark alignments, nor even upon a "gold-standard" alignment of the subset of sequences with known structure. Furthermore, it requires no knowledge of how the MSAs were produced, nor upon how the methods that produced them perform on other sets of sequences. Rather, for each MSA, it first derives, from pairwise correlations among columns, internal evidence of likely 3D contacts among residue positions of the aligned proteins, and then uses the known structures to assess the relative accuracy of this evidence. This approach is based on the principle that, to maintain a protein family's structural fold, interacting residues pairs tend to coevolve, resulting in correlations better seen within accurate alignments. Hence, the degree to which 3D contacts may be correctly inferred from an MSA depends upon its accuracy.

Because eCOMPASS applies to the evaluation of the overall quality of specific sequence alignments that are very large, it cannot be readily evaluated using known benchmark MSAs, nor are we aware of previous approaches to which it can be properly compared. We therefore argue for its validity from its inherent plausibility, its application to simulated gold standard alignments, and its consistency with a completely independent measure of alignment accuracy than the measure eCOMPASS deploys.

We first describe the eCOMPASS algorithm and illustrate its use by applying it to eight pairs of large MSAs obtained from the CDD and PFAM databases and containing a sufficient number of proteins of known structure. We also describe the sort of insights eCOMPASS can reveal regarding the relative quality of such MSAs. Second, we validate it on simulated MSAs generated from realistic Potts models of protein superfamilies versus realignments of the simulated sequences using four different alignment methods. Third, we evaluate its robustness to changes in various hyperparameter settings.

2 Methods

2.1 Input and basic strategy

eCOMPASS takes as input two MSAs of the same set of protein sequences aligned using two different methods. We recommend that the set include at least ten proteins of known structure. The method's basic strategy is, first, to use correlations among columns in each MSA to predict which pairs of columns correspond to residue 3D contacts; and then to check the accuracy of these predictions (measured as described below) using the aligned proteins of known structure. The method assumes that the more accurate the overall MSA, the more accurate will be structural predictions derived from its column correlations. Evidence for the validity of this assumption is provided through analyses of simulated MSAs.

Note that, although eCOMPASS uses a relatively small number of sequences with known structure to vote on the relative accuracy of two MSAs, each structure's vote is based upon evidence derived from all the sequences in each of the MSAs. Thus, an MSA that accurately aligns the structures in question to one another but does a poor job of aligning sequences from a much larger and more diverse protein superfamily, should fare poorly in eCOMPASS's estimation. This contrasts with evaluation methods that use the accurate alignment of a (typically small) test set alone as a proxy for an MSA's more general accuracy. Note also that eCOMPASS requires no "gold standard" alignments whose accuracy must be assumed. It bases its evaluation only on the given MSAs and on the experimentally determined structures.

2.2 Direct Coupling Analysis

In order to infer structural information from correlations between column pairs of each MSA, as a prelude to assessing the accuracy of this information, eCOMPASS first performs on the alignments direct coupling analysis (DCA) (Hopf, et al., 2012; Lunt, et al., 2010; Morcos, et al., 2011; Nugent and Jones, 2012; Weigt, et al., 2009). Residue pairwise correlations were long believed, in principle, to be predictive of structural contacts, but early approaches fell short of expectations due to the confounding effect of indirect correlations: When residues correlate both at positions i and j and at positions j and k, then residues at positions i and k may also correlate even though they fail to interact directly. DCA overcomes this problem by disentangling direct from indirect correlations using a variety of algorithmic strategies. eCOMPASS uses pseudo-likelihood maximum entropy optimization (Marks, et al., 2011; Marks, et al., 2012) as implemented in CCMpred (Seemayer, et al., 2014); this strategy outperformed (Neuwald and Altschul, 2018) DCA programs based either on sparse inverse covariance estimation (Jones, et al., 2012) or on multivariate Gaussian modeling (Baldassi, et al., 2014).

Many multiple alignment methods construct an idealized model to which individual protein sequences are aligned, resulting in some residues being treated as insertions with respect to this model, and therefore left essentially unaligned to residues in other sequences. For an MSA constructed by such a method, it is only the columns corresponding to modeled positions to which we apply DCA, and we effectively ignore all inserted residues. Other multiple alignment methods align all residues in all input sequences, but this usually results in many columns having null characters for most sequences. To apply DCA effectively to such

alignments, we first exclude columns having greater than 50% null characters.

The output of DCA applied to an MSA M_1 is a set K_1 of direct coupling (DC) scores for all of M_1's column pairs. DC-scores correspond to the average product corrected Frobenius norms (Dunn, et al., 2008; Seemayer, et al., 2014). (DCA methods model both one- and two-site statistics, though eCOMPASS makes use of only the later.) We assume only that these scores grow monotonically with the degree of inferred direct coupling between MSA columns. We observe, however, that there is no immediate way to compare the set K_1 with an analogous set K_2 derived from M_2, both because they typically will differ in size, and because there is no clear correspondence between the columns of M_1 and M_2.

We address this issue by using the sequence of each protein with known structure, considered individually, to choose comparable subsets of K_1 and K_2, which we call K_1' and K_2'. Specifically, for a given protein, we first determine the subset R of its residues that are aligned both in a column in M_1 and in a column in M_2. Identifying the residues in R with the MSA columns to which they are aligned, we define K_1' (and K_2' analogously) as the subset of K_1 corresponding to all pairs of residues in R separated by at least m (5 by default) intervening residues within the protein's primary sequence. (We impose this latter condition because we are not interested in predicting close contacts that are imposed by a protein's backbone.) K_1' and K_2' are then of equal size, with elements corresponding to identical pairs of residues within R. Note, however, that each individual structure defines distinct K_1' and K_2', and it is only such sets, constructed from the same structure, that are directly comparable.

2.3 Initial Cluster Analysis

Our approach is based on the assumptions that within a protein family the evolution of structurally interacting residue pairs is likely to be correlated, and that an accurate multiple alignment of sequences in the family should capture information concerning such correlations in the form of high DC scores. Given two MSAs for a protein family, and a particular structure, we have constructed sets of DC scores, K_1' and K_2', each of whose elements correspond to the same set of residue pairs of known 3D distance and are therefore comparable. We note, however, that inherent differences, such as differing numbers of columns, in the MSAs M_1 and M_2 that are used to construct first K_1 and K_2, and then K_1' and K_2', renders problematic the direct comparison of the raw scores within K_1' and K_2'. Instead, we assume only that higher scores within each set should be preferentially associated with closer structural distances.

To measure the strength of the association between DC scores and physical distances, we turn to Initial Cluster Analysis (ICA) (Altschul and Neuwald, 2018). ICA considers an ordered array of L elements, among which D are designated as distinguished, and seeks the initial segment of the array, of length X, with the most surprising number d of distinguished elements, as measured by a p-value. A generalization of ICA that has been applied to DC scores (Neuwald and Altschul, 2018), and which we employ here, adds an ordering to the distinguished elements, and folds into its optimization a statistical measure of the degree to which the higher ranked among the distinguished elements appear earlier in the array. In essence, this generalization can be understood as measuring the degree of congruence between two ordered sets.

Here, we take the array of elements to be the set of DC scores K_1' (or K_2'), ordered from highest to lowest. The distinguished elements are those corresponding to residue pairs whose structural distance is $\leq z$ (with $z = 4$ Å by default). Note that, except for glycine, z is based on the distance between sidechain atoms rather than between α- or β-carbons. ICA returns an S-score (Neuwald and Altschul, 2018) calculated as $S = -\log_{10}(p)$. S-scores have units of log-probability and are therefore directly comparable. Nevertheless, when the relationship between two orderings is known, or strongly suspected, to be significant, an array with a larger number of elements L, and/or a larger number of distinguished elements D, may intrinsically favor the generation of higher or lower S-scores. In such cases, it is best to compare only S-scores generated from arrays with the same L and D. Because the scores S_1 and S_2 we calculate for our two input MSAs from K_1' and K_2' are, by construction, generated using the same L and D, we take their difference $\Delta S = S_1 - S_2$ as a valid measure of the evidence provided by the structure in question for the relative accuracies of MSAs M_1 and M_2. In this study, an S-score can be understood as a statistical measure of the congruence of structural contacts with DC-scores (i.e., average product corrected Frobenius norms).

2.4 Eliminating Structures Likely to be Misaligned

It would be possible to assess the relative quality M_1 and M_2 by evaluating solely how well each MSA aligns the reference structures to one another. However, this would ignore how the vast number of remaining sequences are aligned. In contrast, eCOMPASS measures how well the DC scores derived from each MSA predict 3D-contacts between residue pairs in each reference structure. This assumes, however, that each structure is properly aligned, in the main, within both MSAs, which may not be the case.

To identify reference structures that may be misaligned within a particular MSA, we first determine, for each structure i, the subset R_i of its residues that are aligned by the MSA to residues rather than null characters in all other structures; note that the R_i will be of the same size for all structures. We then compute, for each pair of structures i and j, the quantity $\Delta\mathfrak{D}_{ij}$, defined as the mean, for all pairs of residues a and b within R_i, of the absolute difference between the Cα distance of a to b and the Cα distance within structure j of the residues to which a and b align. It can be seen that $\Delta\mathfrak{D}_{ij} = \Delta\mathfrak{D}_{ji}$, and this quantity may be understood to measure how well sequences i and j are structurally aligned with one another (Hasegawa and Holm, 2009; Holm, et al., 2008). Assuming most structures are on average properly aligned, a structure i that is poorly aligned should have high $\Delta\mathfrak{D}_{ij}$ for most j, and therefore an unusually high mean value of $\Delta\mathfrak{D}_{ij}$ for all $j \neq i$, which we denote as $\Delta\mathfrak{D}_i$. Any structure whose $\Delta\mathfrak{D}_i$ is ≥ 2 SD above the mean is likely to be misaligned and thus to yield unreliable results, and we accordingly may choose to remove it from consideration. We iteratively recalculate until convergence the mean and SD from the remaining $\Delta\mathfrak{D}_i$, and each time remove any structure whose $\Delta\mathfrak{D}_i$ is ≥ 2 SD above the mean. Of course, to apply this approach effectively it is important to have a sufficient number of diverse structures (corresponding by default to proteins sharing $\leq$ 65% sequence identity). After all structures with questionable alignment within either MSA have been removed, we calculate $\overline{\Delta S}$, the mean value of ΔS, both for the remaining structures and for all structures, as two alternative measures of the relative quality of M_1 and M_2.

Note that the number of columns used to calculate the $\Delta\mathfrak{D}_i$ varies from one MSA to another, as of course do the subsets of residues R_i within the various structures. Thus, in contrast to the S_i, the $\Delta\mathfrak{D}_i$ are properly comparable only among different structures for the *same* MSA, but not between one MSA and another. Nevertheless, as we will see below, there is a noticeable tendency for the MSA preferred by the measure $\overline{\Delta S}$ also to yield a lower $\overline{\Delta\mathfrak{D}}$ (mean $\Delta\mathfrak{D}_i$), which can be understood as a rough measure of how well an MSA aligns the reference structures to one another.

2.5 Using simulated Potts model MSAs as gold standards

We created a Potts model for each of 40 CDD/MAPGAPS-generated MSAs (listed in **Table S1**) using CCMpredPy (Vorberg, et al., 2018). To obtain 3D contacts for each Potts model, we created corresponding homology modeled structural coordinates using SWISS-MODEL (Waterhouse, et al., 2018); column pairs corresponding to 3D contacts > 8 Å in the structure are set to zero in the Potts model generated by CCMpredPy. A simulated 5,000 sequence alignment was generated for each Potts model using CCMgen (Vorberg, et al., 2018). We realigned the sequences for each of the simulated MSAs using four different MSA programs (see below) and used eCOMPASS to score each realigned MSA when compared to the corresponding gold standard MSA.

3 Application

3.1 Overview

Most commonly used multiple alignment programs fail to generate plausible MSAs when given as input the numbers of sequences considered in this study, typically in the tens or hundreds of thousands. Therefore, we do not attempt to evaluate these programs, but instead apply eCOMPASS in three ways: (i) to 8 CDD versus PFam MSAs; (ii) to 40 realigned versus gold standard simulated MSAs; and (iii) to 31 CDD versus JackHMMER MSAs using various eCOMPASS hyperparameter settings.

Table 1. Eight pairs of CDD vs Pfam MSAs analyzed here.

name		MSA1		MSA2			avg.
abbr.	# seqs	len	CDD	len	Pfam	#pdb	%id
C2	72,249	102	cd00030	103	PF00168	34	22
CuDX	15,418	110	cd00920	119	PF07732	20	23
HAD	58,031	95	cd01427	95	PF00702	18	21
MBL	70,293	188	cd06262	197	PF00753	32	14
PH	36,099	89	cd00900	105	PF00169	30	17
PTS	9,395	84	cd00133	90	PF02302	13	18
RHOD	61,053	89	cd00158	107	PF00581	33	19
SFTS	35,560	237	cd00016	309	PF00884	21	19
mean:	44,762	124		141		25	19

The numbers of aligned sequences for each domain are given in column 2. Lengths of MSA 1 and 2 are given in columns 3 and 5, respectively, and corresponding CDD and Pfam identifiers are given in columns 4 and 6, respectively. CDD alignments were obtained using, as input to MAPGAPS, the NCBI CDD hierarchical MSA and the sequences present in the corresponding Pfam MSA, as was recently described (Neuwald, et al., 2020). Each Pfam MSAs had been generated automatically by creating a hidden Markov model profile from a Pfam seed alignment and then aligning related sequences to the profile (Sonnhammer, et al., 1998). For each analysis, the number of reference structures and the average % identity shared among aligned regions of known structure are given in columns 7 and 8, respectively.

3.2 CDD versus Pfam MSAs

We illustrate eCOMPASS using 8 pairs of MSAs (**Table 1**), each consisting of one CDD-based MSA (obtained as described in Table 1) and one Pfam MSA (El-Gebali, et al., 2019). These MSA pairs represent the following protein superfamilies: C2 domains (C2); cupredoxins (CuDX); haloacid dehalogenase-like hydrolases (HAD); class B metal β-lactamases (MBL); pleckstrin homology domains (PH); phosphotransferase system subunit IIB (PTS); rhodanese homology domain (RHOD); and sulfatases (SFTS). We obtained a mean of 25 reference structures per domain. Over their domain footprints, on average these share 19% sequence identity, and each structure shares < 50% identity with all other structures. Thus, these represent well the diversity of each superfamily. The eCOMPASS output files are available as Supplementary Material. "CDD" MSAs achieved, on average, higher *S*-scores than Pfam MSAs (**Table 2**). However, because both types of alignments depend on some degree of manual curation, we draw no general conclusion regarding which of these tend to be more accurate. Rather, our aim here is merely to describe eCOMPASS and illustrate its application.

3.3 CDD vs PFam subgroup-specific analyses

Because a protein superfamily is typically composed of multiple families and subfamilies, which may be aligned with differing accuracy, the ΔS scores for different structures should not be considered as drawn from the same underlying distribution and their variance may therefore be very high. Accordingly, when asking which is the more accurate of two MSAs overall, it is better to consider each ΔS score as a separate vote. Assuming independence for simplicity, we calculate the significance of the majority vote using the two-tailed *p*-value for the equiprobable binomial distribution. We expect these *p*-values to correlate to some extent with $\overline{\Delta S}$, the mean ΔS score, but these two quantities may vary considerably in implied significance, or, in principle, even disagree on which is the better MSA. Also, we recognize that even two structures with low sequence identity are not truly independent, so that our calculated *p*-values must be discounted to some extent.

In **Table 2**, we present a summary of eCOMPASS's results for the eight domains considered. After putatively misaligned reference structures are excluded, for four domains (C2, MBL, PH and PTS) eCOMPASS finds unanimity among the remaining structures favoring one of the MSAs. These agreements are statistically significant, with the Pfam MSA favored for C2, and the CDD MSA favored for MBL, PH and PTS. (This frequent unanimity is evidence that the ΔS score is no mere random artifact but is a valid measure for the greater ability of one MSA to encode structural features as directly coupled residue pairs.) For the remaining four domains, neither MSA is preferred with an estimated $p < 0.001$, and the SD of the ΔS values exceeds their absolute mean.

Table 2. eCOMPASS results with outliers excluded.

ID	MSA 1		MSA 2		$\overline{\Delta S}$	SD	$-\log_{10}(p)$
	N_1	$\overline{\Delta \mathfrak{D}}$	N_2	$\overline{\Delta \mathfrak{D}}$			
C2	0	1.12	25	1.15	-9.7	5.3	7.2
CuDX	12	1.19	4	0.89	2.4	5.3	1.1
HAD	4	1.27	10	1.18	-4.3	8.9	0.7
MBL	18	1.74	0	3.28	82.1	16.5	5.1
PH	20	0.99	0	1.43	8.9	6.4	5.7
PTS	12	2.19	0	2.89	11.2	5.6	3.3
RHOD	19	1.57	9	2.18	6.5	10.3	1.1
SFTS	14	1.40	5	1.80	15.5	19.7	1.2

For each domain, values of $\overline{\Delta \mathfrak{D}}$ and $\overline{\Delta S}$ were calculated only after excluding unreliably aligned structures, as described in the text. N_1 and N_2 are the observed number of included structures for which $S_1 > S_2$ and $S_2 > S_1$, respectively. The ΔS-score standard deviation (SD) measures the variability among reference structures for each domain. For the last column, p is calculated as the 2-tail binomial probability for the observed N_1 and N_2, assuming an equal chance for each MSA to have higher ΔS for each structure.

To illustrate and study our procedure for excluding structures, we consider in detail its operation on the PTS domain. In **Table 3**, we show the specific values of S and $\Delta\mathfrak{D}_i$ for each of the 13 reference structures and each MSA. As is apparent, only for structure 3czcA and MSA 1 does $\Delta\mathfrak{D}_i$ exceed the mean by over two SDs, so we exclude this one structure as unreliably aligned. (When the mean and SD for the remaining $\Delta\mathfrak{D}_i$ for MSA 1 are recalculated, no further structures are excluded.) Note that this has the effect of eliminating the one negative ΔS, leaving unanimous preference for MSA 1 among the remaining structures. An examination of the structures eliminated by our procedure for the other seven domains shows that they very often yield outlying values of ΔS, although this is neither expected nor observed to be universally the case.

Table 3. eCOMPASS output for the PTS domain.

pdbid	MSA 1		MSA 2		ΔS	cols	D	L
	S_1	$\Delta\mathfrak{D}_i$	S_2	$\Delta\mathfrak{D}_i$				
3czcA	**29**	2.78	**41**	2.94	**-12.0**	**82**	**100**	**2944**
2wy2D	47	2.10	34	2.60	12.9	77	85	2583
2l2qA	21	2.42	15	3.00	5.9	65	39	1801
4mgeA	51	2.06	38	2.48	12.7	78	93	2659
3nbmA	54	2.24	30	2.87	23.5	76	88	2525
1tvmA	29	2.34	19	2.92	10.2	74	60	2367
5gqsA	31	2.23	15	2.92	15.8	78	79	2647
1vkrA	28	2.12	11	3.07	16.7	71	64	2164
5dleA	32	2.08	24	2.95	7.3	77	97	2590
2r48A	32	2.11	22	2.91	10.1	77	93	2590
4tn5A	24	2.12	16	2.90	7.5	75	86	2453
2kyrA	22	2.37	20	3.07	2.4	77	90	2595
2m1zA	31	2.10	22	2.92	9.0	77	87	2594
mean:		2.24		2.89	9.4			
SD:		0.20		0.17	8.4			

Values for 2czcA are shown in bold to indicate that its $\Delta\mathfrak{D}$ value for MSA 1 is ≥ 2 SD above the mean. The 7th column gives the number of columns shared by MSA 1 and 2 when computing S-scores. Columns 8 and 9 give the values of D and L for the ICA procedure.

Original 3czcA PTS alignment: $\Delta\mathfrak{D}_i$ = 2.78 Å; S_1 = 29.1; ΔS = -12.0

```
3czcA  41 VGE-.AKGLASNYDIVVASNHLI.........HELDGRTNGKLIGLD...NLM 79
2wy2D  41 ETLAgEKG..QNADVVLLGPQIAymlpeiqrlL--PNKPVEVIDSLLy..GKV 87
2l2qA  24 ETRLsEVV..DRFDVVLLAPQSRfnkkrleeiTKPKGIPIEIINTIDy..GTM 72
4mgeA  38 GDAVkTNI..DQADVLLLGPQVRymlssmktlADERNVGIDVINPMHy..GMM 86
3nbmA  80 YGAHyDIM..GVYDLIILAPQVRsyyremkvdAERLGIQIVATRGMEyihLTK 130
```

Corrected 3czcA PTS alignment: $\Delta\mathfrak{D}_i$ = 2.35 Å; S_1 = 41.5; ΔS = +4.3

```
3czcA  41 VGEAkGLASNYDIVVASNHLIhel......DGRTNGKLIGLD---...NLM 79
2wy2D  41 ETLAgEKGQNADVVLLGPQIAymlpeiqrlL--PNKPVEVIDSLLy..GKV 87
2l2qA  24 ETRLsEVVDRFDVVLLAPQSRfnkkrleeiTKPKGIPIEIINTIDy..GTM 72
4mgeA  38 GDAVkTNIDQADVLLLGPQVRymlssmktlADERNVGIDVINPMHy..GMM 86
3nbmA  80 YGAHyDIMGVYDLIILAPQVRsyyremkvdAERLGIQIVATRGMEyihLTK 130
```

Fig. 1. $\Delta\mathfrak{D}_i \geq 2$ SD above the mean for 3czcA is due to misalignment. (top) For the CDD PTS MSA, the sequence corresponding to 3czcA yielded $\Delta\mathfrak{D}_i$ = 2.78 Å, which is 2.7 SD above the mean, suggesting this structure is misaligned relative to the 12 other structures, four of which are shown. As a result, eCOMPASS discarded 3czcA's ΔS value when computing $\overline{\Delta S}$ = 11.2 in Table 2. (bottom) When 3czcA was structurally realigned using Dali (Holm and Rosenstrom, 2010), its $\Delta\mathfrak{D}_i$ decreased to 2.35 Å (1.5 SD above the mean) and its S score increased to 41.5, providing further evidence that it was originally misaligned. The realigned region is highlighted in black; numbers correspond to the residue positions at each end.

It is not eCOMPASS's function to amend the MSAs with which it is supplied. However, to study further the validity of eCOMPASS's procedure for rejecting structures as misaligned, and their corresponding ΔS as unreliable, we used Dali (Holm and Rosenstrom, 2010) to structurally realign 3czcA to the other structures. As shown in **Fig. 1**, given the resulting modified MSA 1, $\Delta\mathfrak{D}_i$ for 3czcA is no longer an outlier, and the ΔS for 3czcA turns positive. Note, however, that sequences closely related to 3czcA in MSA 1 were not realigned; if they had been, presumably the ΔS would have increased further.

One may object to our procedure for excluding a structure, from one or both MSAs, based upon internal evidence that it has been misaligned. Such a structure generally represents not only itself but also the alignment of closely related sequences, and arguably should have a vote equal to that of other structures regarding which alignment is better. In **Table 4** we give the results of our analysis if no structures are excluded. As might be expected, the values of $\overline{\Delta\mathfrak{D}}$ in **Table 4** are higher, although this need not always be the case because the removal of a structure due to a significantly high $\Delta\mathfrak{D}_i$ for one MSA may decrease $\overline{\Delta\mathfrak{D}}$ for the other MSA. Also, for all domains except CuDX, the standard deviation of the ΔS is higher. This too is expected, because, although structures are removed with no reference to ΔS, misaligned structures have a strong tendency to produce outlying values for ΔS, as illustrated, for example, in **Table 3**. Most importantly, however, for all domains the assessment of which is the better MSA is essentially unchanged, by the measure either of $\overline{\Delta S}$ or of the binomial vote N_1 vs. N_2. There appears to be a slight tendency for both $|\overline{\Delta S}|$ and $-\log_{10}(p)$ to decrease with the inclusion of all structures, but this is neither systematic nor coordinated. The advantage of excluding apparently misaligned structures is that this focuses more on the overall quality of the MSAs, as measured by their direct coupling signal, and less on the alignment accuracy of the relatively small number of structures considered. To help assess such distinctions, eCOMPASS computes results using both approaches.

Table 4. eCOMPASS results with outliers included.

ID	MSA 1		MSA 2		$\overline{\Delta S}$	SD	$-\log_{10}(p)$
	N_1	$\overline{\Delta\mathfrak{D}}$	N_2	$\overline{\Delta\mathfrak{D}}$			
C2	3	1.43	31	1.53	-8.1	6.2	6.1
CuDX	15	1.19	5	0.96	2.1	5.1	1.4
HAD	6	1.35	12	1.33	-4.3	8.9	0.6
MBL	31	2.06	1	3.79	74.0	28.0	7.8
PH	29	1.12	2	1.59	9.4	14.3	6.3
PTS	12	2.24	1	2.89	9.4	8.4	2.5
RHOD	24	1.71	9	2.32	6.4	9.6	1.9
SFTS	15	1.52	6	1.84	13.0	24.4	1.1

For some superfamilies neither MSA was significantly favored based on the binomial p-value. For example, for the sulfatases (SFTS) $p = 0.06$ and, among the retained ΔS scores, 14 were positive (favoring MSA 1) and 5 were negative (favoring MSA 2). The variability in ΔS scores was very high with a SD of 19.7 and a mean of 15.5. Similar results were obtained when using all ΔS scores. This suggests that MSA 1 better aligns some functionally divergent subgroups while MSA 2 better aligns others. This may occur, for example, when an MSA is generated by a query-based iterative alignment method, such as PSI-BLAST (Altschul, et al., 1997) or JackHMMER (Johnson, et al., 2010), resulting in subgroups closely

related to the query being more accurately aligned than distantly-related subgroups. The Pfam MSAs used for this study were generated using a similar profile-based alignment method.

By providing a more articulated description of relative alignment quality than would a single measure of overall quality, eCOMPASS may aid the curation of hierarchical MSAs (Yang, et al., 2020), which were provided as input to MAPGAPS to generate the MSA 1 alignments used here. For instance, for the SFTS domain, the structure 4uplA, which is a member of the phosphonate monoester hydrolase family (i.e., cd16028), has the lowest ΔS score (-53.2) and the highest $\Delta\mathfrak{D}_I$ (2.99 Å) for MSA 1 (see Supplementary Data). This suggests that, by further curating the cd16028 subgroup, one could improve the CDD hierarchical MSA and thus the SFTS MSA generated from it.

Finally, as discussed above, $\overline{\Delta\mathfrak{D}}$ scores should only be compared with caution because both the numbers and the nature of the residue pairs used to compute Cα–Cα distances differ between MSAs. For example, unlike other domains, the C2 MSA deemed superior by the measure of $\overline{\Delta S}$ (Tables 2 and 4) yielded higher $\overline{\Delta\mathfrak{D}}$. This illustrates how relying on $\overline{\Delta\mathfrak{D}}$ scores may miss distinctions between MSAs revealed by better justified and statistically based $\overline{\Delta S}$ scores.

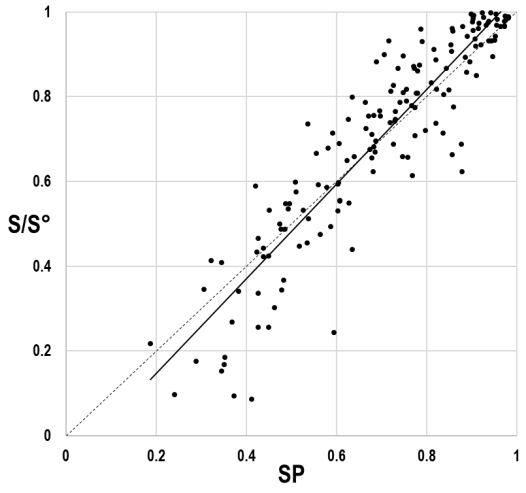

Fig. 2. ***S/S°* as a function of SP-score for simulated gold standard versus realigned MSAs.** The 160 data points represent 40 simulated (gold standard) MSAs, each of which is compared to 4 different realigned MSAs of the corresponding simulated sequences. The solid line corresponds to the regression line and the dotted line to $y = x$.

3.4 Program-aligned vs gold standard simulated MSAs

Using the procedure described in Methods, we created 40 simulated gold standard MSAs, each with a single associated structure. We realigned the sequences of each MSA using four programs: GISMO (v3.1) (Neuwald and Altschul, 2016), Kalign 3 (Lassmann, 2020), MAFFT (v7.471) (Katoh and Standley, 2014), and MUSCLE (v3.7) (Edgar, 2004). To compute each realigned MSA's distance from its associated gold standard, we calculated an SP-score (from "Sum of the Pairs"), which is the proportion of aligned pairs of residues within the gold standard that are aligned identically within the realigned MSA. We then used eCOMPASS to compare each realigned MSA to its corresponding gold standard MSA. As described above, given two MSAs eCOMPASS generates directly comparable scores, which we here denote as S for the realigned MSA and as $S°$ for the gold standard MSA. Notably, as expected, in all cases the S-score is less than the $S°$-score. To study how well the relative values of S and $S°$ correspond to the distance between the realigned and gold standard MSAs, we plot in **Fig. 2** $S/S°$ versus SP for each case. There is clearly a strong and close to linear correlation between $S/S°$ and SP, with the Pearson correlation coefficient equal to 0.92. The regression line has a slope of 1.117 and a y intercept of -0.077, suggesting that $S/S°$ is a good and relatively direct proxy for gold standard distance. Hence, for real protein sequence alignments, where we do not have gold standards for comparison, we may use comparable S scores as proxies for alignment accuracy.

3.5 CDD vs JackHMMER MSA analyses

To further explore the utility and robustness of eCOMPASS, we compared the 40 CDD MSAs, upon which our simulated MSAs were based, to corresponding MSAs aligned with JackHMMER (JHM) (Johnson, et al., 2010) using an arbitrary sequence as the query (**Table S1**). To reduce sequence redundancy, we removed from each MSA all but one sequence among those sharing ≥95% sequence identity using either cd-hit (Fu, et al., 2012) or PurgeMSA (Neuwald, et al., 2020). Note that this analysis allows the inclusion of more reference structures because, unlike the CDD vs Pfam analysis, the number of structures included was not predefined by Pfam. To identify domains for which a clearly significant distinction was at least possible, we focused on 31 of the 40 domains having at least 18 distinct structures, which could, in principle, yield a two tailed binomial probability $p < 10^{-5}$. Among these the CDD MSA was significantly better at the $p < 10^{-3}$ level for 12 domains whereas the JHM MSA was significantly better for 6 domains (**Fig. 3**).

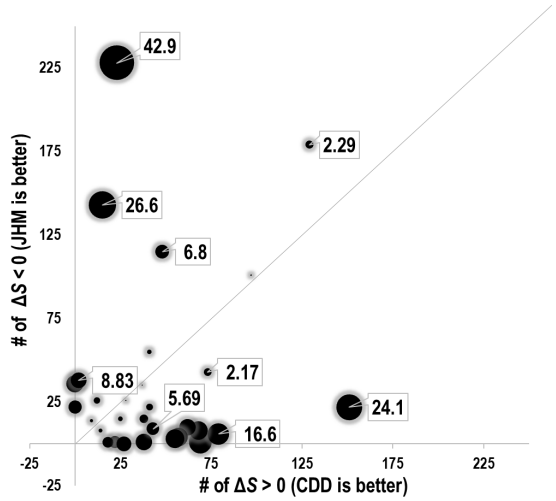

Fig. 3. eCOMPASS analysis of CDD vs Jackhmmer (JHM) MSAs. Data points represent 31 comparisons with the x- and y-axes corresponding to the numbers of reference structures for which $\Delta S > 0$ and $\Delta S < 0$, respectively. Hence, data points below and above the diagonal line correspond to analyses favoring the CDD and JHM MSA, respectively. The area of each bubble is proportional to $-\log_{10}(p)$, the values of which are indicated for several data points.

To evaluate the robustness of eCOMPASS, we reran each of these analyses using various CCMpred hyperparameter settings. (Another variable is the DCA implementation used, which, however, is too technically challenging to investigate here.) Using either flat (uniform) priors or Jeffreys uninformed priors [28] yielded essentially identical results (Fig. S1). We also ran eCOMPASS with maximum residue pair 3D contact cutoffs of 4, 5, and 6 Å (**Fig. 4**), with alternative CCMpred

sequence reweighting thresholds of 70, 80, and 90% (**Fig. 5 top**), and with L1 regularization strengths of 0.1, 0.2, and 0.3 (**Fig. 5 bottom**). Notably, in only one case did two different parameter settings yield conflicting results both at a significance level ≤ 0.01. This arose for the L1 regularization parameter and the AAT_1 domain, for which conflicting results were reported with *p*–values of 0.005 and 0.002.

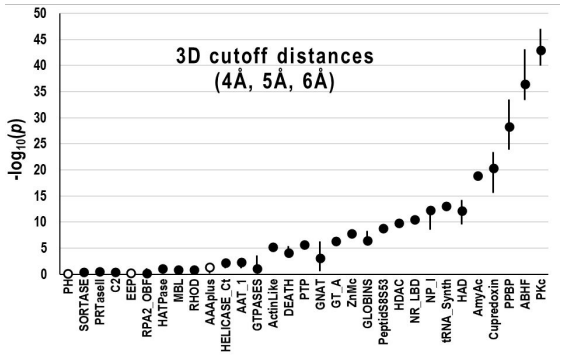

Fig. 4. Influence of the 3D contact cutoff on eCOMPASS results. Plots indicate probabilities for CDD MSAs versus JHM MSAs using 4, 5 and 6 Å cutoffs. Circles correspond to median values and vertical lines to the high and low values. Closed or open circles indicate that the MSAs considered better are consistent or inconsistent, respectively, across the 3 settings. Domains are ordered left to right by the maximum of their three $-\log_{10}(p)$ values.

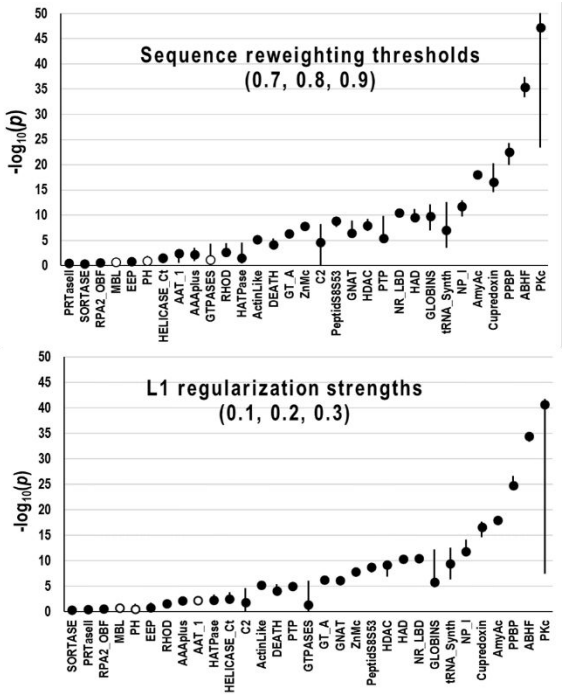

Fig. 5. Influence of DCA hyperparameter settings on results. Plots indicate probabilities for CDD MSAs versus JHM MSAs using the 3 settings indicated. Circles correspond to median values and vertical lines to the high and low values. Closed or open circles indicate that the MSAs considered better are consistent or inconsistent, respectively, across the 3 settings. (top) CCMpred reweighting thresholds. (bottom) CCMpred L1 regularization strengths.

The observed variability in the binomial probability yielded by different parameter settings is likely due to changes the implicit nature of the MSAs, of the ICA array or of both. For example, decreasing the CCMpred reweighting threshold (Seemayer, et al., 2014) is likely to decrease the DCA signal from highly populated subgroups.

4 Discussion

eCOMPASS computes a statistical score ($\overline{\Delta S}$) that compares the accuracy of two large MSAs and that is based on all the aligned sequences and on a set of reference structures. This score exploits the DC-signal implicit in each alignment and whose strength presumably depends on the degree to which homologous residues are accurately aligned. eCOMPASS's strategy constitutes a departure from current approaches. These typically rely upon a benchmark set, consisting of a small number of sequences aligned using structural data. However, they are essentially blind to the alignment accuracy of sequences absent from the set. Unlike other programs for assessing MSA quality (Ahola, et al., 2008; Lassmann and Sonnhammer, 2005; O'Sullivan, et al., 2003; Pei and Grishin, 2001; Song, et al., 2006; Thompson, et al., 2001), eCOMPASS provides measures of statistical significance, can handle extremely large MSAs, requires neither a gold standard MSA nor a structural alignment, and can assess the alignment quality of subgroups within an MSA.

Almost all multiple alignment construction methods employ some objective function of alignment quality which they attempt to optimize. For assessing the relative accuracy of two multiple alignments, relying upon the objective function used for either's construction will of course bias the results, so it is best to seek an independent measure. The congruence of structural contacts with alignment-derived DCA scores provides a convenient such measure, and one that avoids reliance upon a set of gold standard alignments.

Several recent multiple alignment construction methods (Muntoni, et al., 2020; Talibart and Coste, 2020; Talibart and Coste, 2020; Wilburn and Eddy, 2020) incorporate DCA models into the objective functions they seek to optimize. To the extent that these models have been derived from particular structures, applying eCOMPASS to their evaluation using these very structures is likely to bias eCOMPASS's results in favor of the resulting multiple alignments. How to extend eCOMPASS to the comparison of such multiple alignments, or at least how to mitigate any confounding effects, is a question for further research. In this paper, however, none of the alignments of real proteins studied here were constructed with the use of a Potts model.

Recently, Muntoni et al. (2020), in comparing the alignments constructed by their program DCAalign to those produced by other programs, used one method very similar in spirit to that of eCOMPASS. From alignment-derived pairwise coupling scores, they predicted contacting residue pairs and then, with reference to a known structure, plotted the true positive prediction rate as a function of the number of predictions made. It should be possible to derive from the resulting graphs a statistically-based measure, similar to our ΔS, for the relative accuracy of the two alignments. Following, for example, the approach of (Schaffer, et al., 2001), one could calculate a ROC (receiver operating characteristic) score from a variant of each graph, and then infer *p*-values for the difference of these scores. Whether such a statistical approach is superior to the one taken here is an avenue for further study.

Ideally, eCOMPASS should be applied using a set of reference structures representing diverse subgroups within a superfamily, as in the examples here. Then, in addition to providing an assessment of overall

alignment accuracy, eCOMPASS can identify those subgroups that are least accurately aligned, as an aid to improving MSA methods. This raises the issue of multiple conformations for the same protein, which is a major concern for DCA. A future version of eCOMPASS might provide the option of choosing the highest DC-score among alternative conformations for each residue pair. In order to investigate directly coupled residue pairs corresponding to a subgroup specific conformation, such as we reported recently (Tondnevis, et al., 2020) , it may be useful to apply eCOMPASS to subgroup alignments within a superfamily MSA.

For MSA methods that fail to incorporate information from DCA into their objective functions, the statistical significance of the agreement between DC-scores and 3D contacts within available structures serves as a measure of alignment accuracy that is independent of the criteria used in constructing the MSA. In any case, eCOMPASS should be uniquely useful for evaluating the extremely large MSAs typically required for deep learning protein sequence analyses and for statistical analyses requiring a vast amount of sequence data.

Acknowledgements

We thank Christopher Lanczycki for critical reading of the manuscript.

Funding

This work was supported by a National Institute of General Medical Sciences grant [R01 GM125878 to A.F.N.] and by a National Science Foundation grant [BIO MCB 1817942 to B.D.K.]. S.F.A. is supported by the intramural research program of the National Institutes of Health, National Library of Medicine.

Conflict of Interest: none declared.

References

Ahola, V., *et al.* Model-based prediction of sequence alignment quality. *Bioinformatics* 2008;24(19):2165-2171.

Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25(17):3389-3402.

Altschul, S.F. and Neuwald, A.F. Initial Cluster Analysis. *J Comput Biol* 2018;25(2):121-129.

Ashkenazy, H., *et al.* Multiple Sequence Alignment Averaging Improves Phylogeny Reconstruction. *Syst Biol* 2019;68(1):117-130.

Baldassi, C., *et al.* Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS One* 2014;9(3):e92721.

Dunn, S.D., Wahl, L.M. and Gloor, G.B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 2008;24(3):333-340.

Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004;5(1):113.

Edgar, R.C. Quality measures for protein alignment benchmarks. *Nucleic Acids Res* 2010;38(7):2145-2153.

El-Gebali, S., *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;47(D1):D427-D432.

Fletcher, W. and Yang, Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol* 2010;27(10):2257-2267.

Fu, L., *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28(23):3150-3152.

Hasegawa, H. and Holm, L. Advances and pitfalls of protein structural alignment. *Curr Opin Struct Biol* 2009;19(3):341-348.

Holm, L., *et al.* Searching protein structure databases with DaliLite v.3. *Bioinformatics* 2008;24(23):2780-2781.

Holm, L. and Rosenstrom, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res* 2010;38(Web Server issue):W545-549.

Hopf, T.A., *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012;149(7):1607-1621.

Johnson, L.S., Eddy, S.R. and Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 2010;11:431.

Jones, D.T., *et al.* PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 2012;28(2):184-190.

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

Kim, C. and Lee, B. Accuracy of structure-based sequence alignment of automatic methods. *BMC Bioinformatics* 2007;8:355.

Lassmann, T. Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics* 2020;36(6):1928-1929.

Lassmann, T. and Sonnhammer, E.L. Automatic assessment of alignment quality. *Nucleic Acids Res* 2005;33(22):7120-7128.

Levy Karin, E., Susko, E. and Pupko, T. Alignment errors strongly impact likelihood-based tests for comparing topologies. *Mol Biol Evol* 2014;31(11):3057-3067.

Lunt, B., *et al.* Inference of direct residue contacts in two-component signaling. *Methods Enzymol* 2010;471:17-41.

Marks, D.S., *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 2011;6(12):e28766.

Marks, D.S., Hopf, T.A. and Sander, C. Protein structure prediction from sequence variation. *Nat Biotechnol* 2012;30(11):1072-1080.

Morcos, F., *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 2011;108(49):E1293-1301.

Muntoni, A.P., *et al.* Aligning biological sequences by exploiting residue conservation and coevolution. *Phys. Rev. E* 2020;102(6):062409.

Neuwald, A.F. Rapid detection, classification and accurate alignment of up to a million or more related protein sequences *Bioinformatics* 2009;25(15):1869-1875.

Neuwald, A.F. Protein domain hierarchy Gibbs sampling strategies. *Statistical applications in genetics and molecular biology* 2014;13(4):497-517.

Neuwald, A.F. and Altschul, S.F. Bayesian Top-Down Protein Sequence Alignment with Inferred Position-Specific Gap Penalties. *PLoS Comput Biol* 2016;12(5):e1004936.

Neuwald, A.F. and Altschul, S.F. Statistical investigations of protein residue direct couplings. *PLoS Comput Biol* 2018;14(12):e1006237.

Neuwald, A.F., Aravind, L. and Altschul, S.F. Inferring joint sequence-structural determinants of protein functional specificity. *Elife* 2018;7.

Neuwald, A.F. and Hirano, T. HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. *Genome Research* 2000;10(10):1445-1452.

Neuwald, A.F., *et al.* Obtaining extremely large and accurate protein multiple sequence alignments from curated hierarchical alignments. *Database (Oxford)* 2020;2020.

Neuwald, A.F., Lanczycki, C.J. and Marchler-Bauer, A. Automated hierarchical classification of protein domain subfamilies based on functionally-divergent residue signatures. *BMC Bioinformatics* 2012;13(1):144.

Neuwald, A.F. and Poleksic, A. PSI-BLAST searches using hidden markov models of structural repeats: prediction of an unusual sliding DNA clamp and of beta-propellers in UV-damaged DNA-binding protein. *Nucleic Acids Res* 2000;28(18):3570-3580.

Nugent, T. and Jones, D.T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci U S A* 2012;109(24):E1540-1547.

O'Sullivan, O.*, et al.* APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics* 2003;19 Suppl 1:i215-221.

Pei, J. and Grishin, N.V. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 2001;17(8):700-712.

Schaffer, A.A.*, et al.* Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* 2001;29(14):2994-3005.

Seemayer, S., Gruber, M. and Soding, J. CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* 2014;30(21):3128-3130.

Song, B.*, et al.* ARCS: an aggregated related column scoring scheme for aligned sequences. *Bioinformatics* 2006;22(19):2326-2332.

Sonnhammer, E.L.L.*, et al.* Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research* 1998;26(1):320-322.

Talibart, H. and Coste, F. ComPotts: Optimal alignment of coevolutionary models for protein sequences. *bioRxiv* 2020:2020.2006.2012.147702.

Talibart, H. and Coste, F. PPalign: Optimal alignment of Potts models representing proteins with direct coupling information. *bioRxiv* 2020:2020.2012.2001.406504.

Thompson, J.D.*, et al.* A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* 2011;6(3):e18093.

Thompson, J.D.*, et al.* Towards a reliable objective function for multiple sequence alignments. *J Mol Biol* 2001;314(4):937-951.

Tondnevis, F.*, et al.* Deep Analysis of Residue Constraints (DARC): identifying determinants of protein functional specificity. *Sci Rep* 2020;10(1):1691.

Toshchakov, V.Y. and Neuwald, A.F. A survey of TIR domain sequence and structure divergence. *Immunogenetics* 2020;72(3):181-203.

Vorberg, S., Seemayer, S. and Soding, J. Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. *PLoS Comput Biol* 2018;14(11):e1006526.

Waterhouse, A.*, et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018;46(W1):W296-W303.

Weigt, M.*, et al.* Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009;106(1):67-72.

Wilburn, G.W. and Eddy, S.R. Remote homology search with hidden Potts models. *PLoS Comput Biol* 2020;16(11):e1008085.

Yang, M.*, et al.* NCBI's Conserved Domain Database and Tools for Protein Domain Analysis. *Curr Protoc Bioinformatics* 2020;69(1):e90.