Statistical evidence for underlying protein molecular properties

Statistical clues regarding underlying protein molecular properties
Statistical evidence for underlying protein molecular properties
Decomposing the constraints imposed on proteins by underlying molecular properties
Decomposing the molecular constraints imposed on protein sequence and structure
Decomposing the sequence and structural constraints due to protein molecular properties
Characterizing the functional constraints associated with protein molecular properties
Decomposing the functional constraints imposed on protein sequence and structure

Andrew F. Neuwald

Institute for Genome Sciences and Department of Biochemistry & Molecular Biology,
University of Maryland School of Medicine, HSF-I, Room 134, 20 Penn Street, Baltimore, MD 21201
Tel: 410-706-6724; Fax: 410-706-1482
E-mail: aneuwald@som.umaryland.edu

Contents

1. Introduction

Knowing a protein's sequence and structure fails to reveal its biochemical and cellular function nor does it reveal the underlying molecular mechanisms responsible for that function. Such mechanisms are mediated by a protein's chemical properties within the context of its molecular environment. Therefore, a complete understanding of how protein molecular machines work requires a detailed, atomic level analysis that also includes molecules with which it functionally interacts. Mutational analysis is often used to study a protein's function determining residues. However, a drawback to this approach is that mutation of one residue can propagate structural and chemical perturbations to other regions of the protein—thereby confounding interpretation of the potential role of the residue being mutated. Moreover, certain functional properties may be mediated by multiple residues, so that assigning specific functional roles to individual residues may not be feasible. Molecular dynamics simulations provide some level of insight into protein properties, but are limited by the need to rely on classical physics; inasmuch as a protein's molecular properties are ultimately quantum mechanical, a proper understanding of how protein molecular machines work would require a detailed quantum mechanical analysis at the atomic level, which, together with these other considerations, puts such an understanding well out of reach. This situation is like that confronted and addressed by the founders of statistical mechanics, such as Ludwig Boltzmann, James Clerk Maxwell, and Josiah Willard Gibbs: Because modeling the behavior of each individual molecule within a large ensemble is intractable, they instead chose to model a molecular system statistically. Here we take a similar approach as an initial step toward understanding currently unknown protein molecular properties. This involves seeking evidence for underlying properties through statistical analyses of large numbers of protein sequences and structures—interpreted in the light of published experimental studies. This requires identification of shared and divergent sequence and structural patterns that may be quite subtle and thus not at all obvious in the absence of measures of statistical significance.

To motivate this topic, consider how two proteins lacking significant pairwise sequence similarity and performing very different functions may appear, through a deeper analysis, to share a common structure core and catalytic mechanism. **Fig. 1.1** illustrates this for soluble chloroperoxidase (CPO) and integral membrane associated type 2 phosphatidic acid phosphatase (PAP2). CPO catalyzes the production of hypochlorous acid (i.e., bleach), which is produced by a parasitic fungus to break down the host cell wall. Membrane-associated phosphatidic acid phosphatase (PAP), on the other hand, is involved in lipid metabolism and signal transduction by catalyzing the dephosphorylation of phosphatidic acid to produce diacylglycerol and orthophosphate. **Fig. 1.1A** shows the subtle sequence similarity that these enzymes share and **Fig. 1.1B,C** proposes a common catalytic core. The CPO reaction is facilitated by a vanadate ion bound to the active site. Because vanadate is a transition state analog of phosphate, this structural relationship suggests a catalytic mechanism for PAP2 (**Fig. 1.1D**). It also raises the question: Which residues within CPO and PAP2 are responsible for their distinct functions? Addressing such questions lies at the heart of our approach.

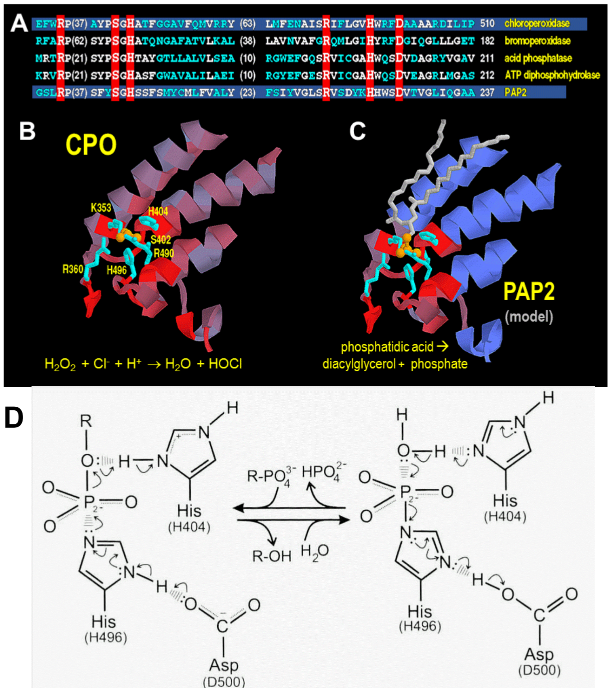

A
EFWRP(37) AYPSGHATFGGAVFQMVRRY (63) LMFENAISRIFLGVHWRFDAAAARDILIP 510 chloroperoxidase
RFARP(62) SYPSGHATQNGAFATVLKAL (38) LAVNVAFGRQMLGIHYRFDGIQGLLLGET 182 bromoperoxidase
MRTRP(21) SYPSGHTAYGTLLALVLSEA (10) RGWEFGQSRVICGAHWQSDVDAGRYVGAV 211 acid phosphatase
KRVRP(21) SYPSGHASFGWAVALILAEI (10) RGYEFGESRVICGAHWQSDVEAGRLMGAS 212 ATP diphosphohydrolase
GSLRP(37) SFYSGHSSFSMYCMLFVALY (23) FSIYVGLSRVSDYKHHWSDVTVGLIQGAA 237 PAP2
B
CPO
K353
H404
S402
R490
R360
H496
$H_2O_2 + Cl^- + H^+ \rightarrow H_2O + HOCl$
C
PAP2
(model)
phosphatidic acid →
diacylglycerol + phosphate
D
R
H
His
(H404)
R-PO_4^{3-}
HPO_4^{2-}
R-OH
H_2O
His
(H496)
Asp
(D500)
His
(H404)
His
(H496)
Asp
(D500)

Fig. 1.1. Structural relationship between soluble haloperoxidases and integral membrane phosphatases [13]. **A**. Homologous regions within haloperoxidases and phosphatidic acid phosphatases (PAPs) identified using the Gibbs sampler described below. Catalytic residues are highlighted in red. **B-C**. Coloring scheme: vanadate and transition state phosphate, *orange*; residue sidechains interacting with vanadate or phosphate, *cyan*. **B**. The vanadium binding domain of chloroperoxidase (CPO) (pdb: IVNC). The vanadate ion (orange), which is a transition state analog of phosphate, is bound from below by His-496 and from above by azide (not shown). The backbone is shaded from gray to red proportional to the increasing degree of sequence conservation. **C**. Model of the active site of mouse PAP2 based on the CPO structure. Predicted trans-membrane regions in PAP2 are shown in blue and the diacylglycerol moiety of phosphatidic acid in light gray. **D**. Proposed catalytic mechanism for PAP2-related phosphatases. Residue positions in parentheses indicate corresponding active site residues within CPO.

In *Novum Organum* [2], Francis Bacon described the scientific method as involving: (i) the compilation of observational data, (ii) followed by the categorization of these observations and the generation of hypotheses, (iii) from which may then follow the accumulation of additional empirical results through further experimentation. Our approach facilitates the second step by characterizing statistically significant constraints imposed on protein sequence and structural data. Our *core hypothesis* is that the most statistically significant constraints reflect protein properties responsible for underlying molecular mechanisms. Because we focus on properties that biochemical and structural studies have thus far failed to identify, we allow the data itself to reveal its most statistically surprising features without making assumptions about what should be found. We argue that, in the absence of relevant biochemical studies, it is only possible to directly link individual residues to other residues, and such residue sets to structural features. Hence, our approach focuses on characterizing these (observed) properties rather than on directly predicting (unobserved) biochemical properties. Augmenting visualization of functionally imposed constraints in this way with a knowledge of biochemical, cellular, and structural properties can lead to plausible hypotheses for experimental design.

This approach is applied to protein superfamilies that have diverged into subgroups, each of which fills a functional niche compatible with the superfamily's common "core" structure. Within each subgroup, proteins from distinct phyla often conserve residues at (non-active site) sequence positions that proteins in other related subgroups fail to conserve. Often a subgroup, *G*, is composed of smaller subgroups, each of which conserves both residues due to *G*'s functional constraints and other residues due to constraints imposed by its own, more specialized function. Repeated rounds of this evolutionary process have led to hierarchically interrelated patterns of correlated residues and, in some cases, to networks of functionally critical residues embedded within structurally defined clusters. Moreover, for proteins sharing a common core structure, 3D contacts between pairs of (often non-conserved) residues generally produce correlated substitution patterns: Over evolutionary time substitutions at one residue position often result in compensating substitutions at other positions to maintain critical interactions. This leads to a correspondence between covarying residue pairs and 3D contacts—a property utilized by structure prediction methods, such as AlphaFold. Other constraints are associated with often subtle internal repeats within protein domains. Hence, our approach looks for statistically significant constraints appearing as residues co-conserved in functionally related subgroups, as subtle pairwise correlations and internal repeats, and as correlations among these sequence features or with structural features. Our aim is to obtain statistical evidence of underlying protein properties required for biological function. Such an analysis can also be applied, of course, to proteins that functionally interact with proteins of primary interest.

Residues associated with functional divergence. Presumably residues that are responsible for the functional specificity of a given subgroup of protein superfamily are conserved within that subgroup but not within closely related subgroups. Consider, for example, Ran GTPase, which belongs to the P-loop GTPase superfamily and which is essential for the translocation of RNA and proteins through the nuclear pore complex. A sequence alignment of Ran GTPases reveals a very high degree of conservation across diverse eukaryotic phyla and kingdoms (**Fig. 1.2**). Most of these conserved residues are not directly involved in catalysis and differ from the residues conserved in other P-loop GTPases. At some conserved positions, even minor side-chain modifications—such as substitution of leucine for isoleucine, which merely involves rearrangement of a methyl group—appear to be consistently eliminated by natural selection. Hence, such residues may establish interactions with precise geometric and/or chemical constraints important for Ran specific functions.

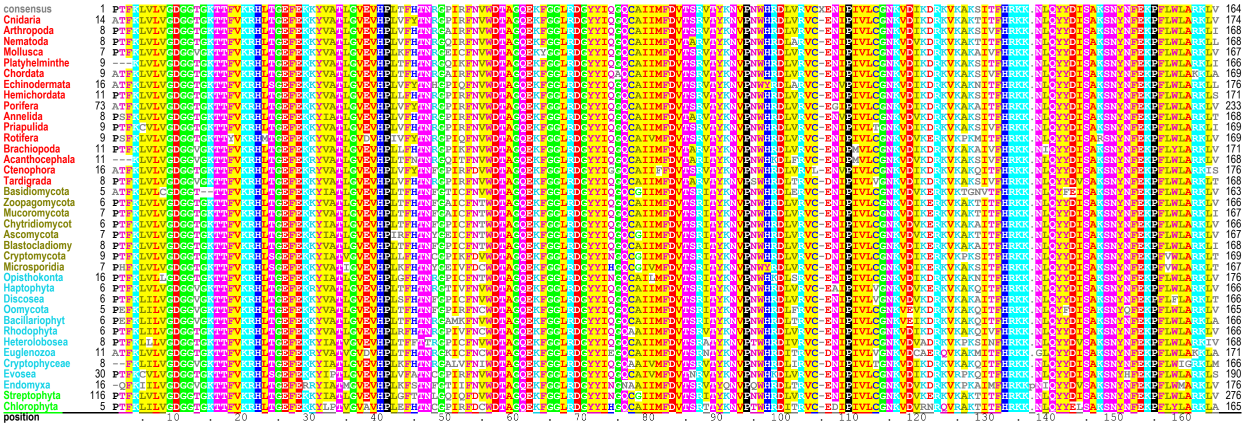
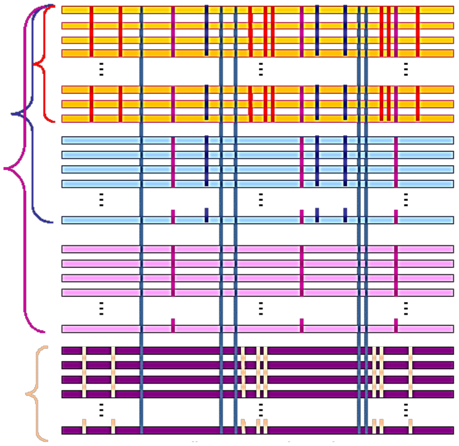


Fig. 1.2. Ran family sequence alignment. The leftmost column specifies each sequence's phylum; these are colored by eukaryotic 'kingdoms', as follows: metazoans, *red*; fungi, *dark yellow*; protozoans, *cyan*; plants, *green*. Biochemically similar conserved residues are colored similarly.



Moreover, within a superfamily alignment, such as that of all P-loop GTPases, there are residues specifically conserved within each protein subgroup as well as residues conserved across multiple subgroups; this is illustrated schematically in the figure on the right, where similarly colored horizontal bars represent sequences corresponding to subgroups, vertical bars represent conserved residue positions, and brackets denote hierarchically-arranged subgroups of families sharing conserved residues that presumably are responsible for structural or mechanistic similarities or both. For P-loop GTPases there are dozens of subgroups that may be hierarchically aligned and analyzed in this way. This construct is termed a hierarchical multiple sequence alignment (hiMSA).

Identifying diverse types of statistically significant constraints within sequence (and structural) data can suggest plausible hypotheses regarding the roles of various categories of function determining residues. Such residues may be remote from an enzyme's active site and may often mediate function through dynamic interactions with each other or with specific cofactors and substrates. For this reason, it may be helpful to apply such an analysis in conjunction with molecular dynamics simulations. Characterizing protein properties in this way can help guide protein engineering efforts, provide insights into the molecular basis of human disease, identify potential antibiotic target sites in bacterial proteins, and aid the design of target-specific drugs.

Amenability to analysis. Proteins belonging to a large, functionally, and taxonomically diverse superfamily is more likely to yield statistically significant, interpretable results than is a less diverse, smaller set of proteins. It is important, of course, to accurately align the sequences. Rapidly evolving proteins, as is typical for viral proteins, may not be amenable to analysis due to a lack of subgroup-specific conserved residues. Distinguishing between conservation due to functional constraints and that due to recent common descent is not possible for closely related species. For this reason, it is best to analyze patterns conserved across distinct phyla or classes. Keep in mind that multiple sequence alignments may contain pseudogene products or sequencing errors—though with sufficient data this should merely add a small amount of noise to an otherwise strong signal.

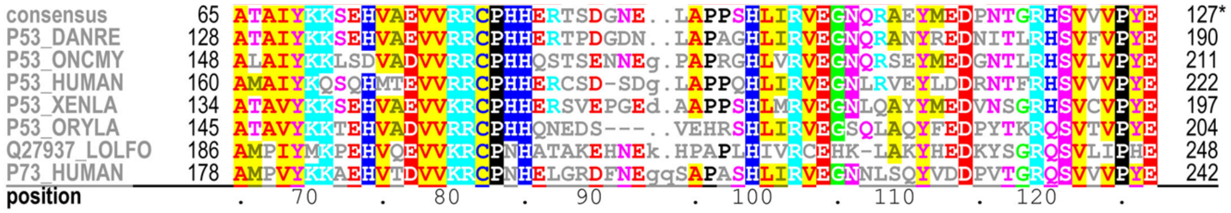
Interpretation of results. An analysis provides functional clues but tells us nothing about the specific biochemical nature of those functions. Hence, biochemical interpretation of an analysis is the hardest part and for certain constraints may be impossible at any given time. Nevertheless, by characterizing selective constraints imposed on proteins over long periods of evolutionary time, an analysis aids formulation of experimentally testable hypotheses regarding protein properties. To best do this, it is helpful to examine constraints in consideration of available protein structures, ideally in complex with functionally interacting cofactors and/or other cellular components, along with other biochemical, and biophysical data. Keep in mind that there may be a lot of currently unknown functions associated with the observed residue constraints. These may include structural interactions, interactions with other cellular components required for cleavage or binding sites or for downstream signaling, and residues required for catalysis, protein folding, or intracellular transport. So, one should avoid jumping to conclusions regarding what functions constrained residues might perform.

Testing hypotheses. Testing of hypotheses generated by an analysis typically involves observing the biochemical phenotype of a protein after mutating one or more pattern residues. Mutant proteins should be checked (e.g., using Circular dichroism) to see whether its structure is disrupted by the mutation, which presumably could disrupt all biochemical activity. If so, then no specific function can be assigned to such residues. When a network of conserved residues is associated with a certain function there may be only a slight degradation in that function when only one such residue is mutated. Ideally, we would like to observe a striking phenotype for the mildest mutation possible. We show such an example below for a bacterial clamp loader subunit where mutation of a threonine residue to valine (involving merely changing a sidechain -OH to -CH_3) leads to a dramatic change in DNA clamp loader function.

Outline. **Chapter 2** provides a brief description of statistics and probability using Bayesian and frequentist methods and of **Markov chain Monte Carlo (MCMC)** sampling. **Chapter 3**, shows how MCMC sampling is used to construct a **multiple sequence alignment (MSA)** of subtle internal repeats, and of full-length protein domains while assessing the advantages of MCMC sampling over other methods. It also illustrates the detection and gapped alignment of subtle internal repeats. MSAs are required as a starting point for many types of sequence/structural analyses, including the statistical characterization of constraints associated with protein functional divergence. **Chapter 4** introduces the notion of a hierarchical (hi)MSA, which arranges an MSA into evolutionarily related divergent subgroups; it also describes a search procedure that uses a hiMSA to detect and multiply align very large numbers of related protein sequences more accurately than otherwise possible. **Chapter 5** shows how to construct a hiMSA using Bayesian Partitioning with Pattern Selection (BPPS), which applies MCMC sampling to separate aligned sequences into divergent subgroups based on constraints distinguishing each subgroup from closely related subgroups. This identifies putative function determining residues that, in consideration of available structural and biochemical information, reveals relationships between protein sequence divergence and functional specificity. **Chapter 6** describes how to identify significant correspondence between sequence and structural constraints—including **direct coupling analysis (DCA)**, which is used to predict structurally interacting residues based on pairwise correlations within an MSA. **Chapter 7** shows how to perform the preceding analyses in a query-specific manner and thereby speed up an analysis when the focus is on a specific protein. **Chapter 8** shows how to sort through vast amounts of sequence and structural data to identify the most statistically striking residue constraints. **Chapter 9** describes various auxiliary programs used for the previously described analyses. **Chapter 10** illustrates a complete protein sequence/structural analysis using our suite of programs. **Appendices 1-6** describe the underlying statistical models upon which our approach is based. **Appendix 7** shows a set of contrast alignments for DNA clamp loader AAA+ subunits which illustrates how constrained residues are visualized.

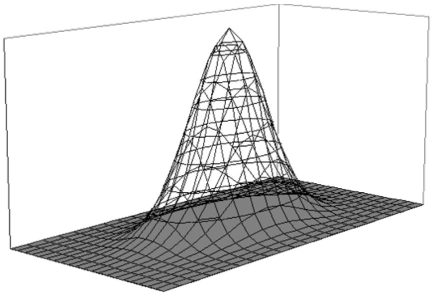
AI versus conventional methods. Some have argued that AI methods have made most conventional computational and statistical methods obsolete. However, unless protein functional determinants are characterized first and then used for training, it is unclear how AI might be used to decompose the sequence constraints imposed on protein sequences and thereby characterize functional determinants. At some point, of course, such AI methods might be developed. Note that AI programs, such as AlphaFold, currently require as input an MSA and a DCA analysis of the MSA.

2. Probabilistic and statistical modeling of proteins



```
consensus      65 ATAIYKKSEHVAEVVRRCPHHERTSDGNE..LAPPSHLIRVEGNQRAEYMEDPNTGRHSVVVPYE 127*
P53_DANRE     128 ATAIYKKSEHVAEVVRRCPHHERTPDGDN..LAPAGHLIRVEGNQRANYREDNITLRHSVFVPYE 190
P53_ONCMY     148 ALAIYKKLSDVADVVRRCPHHQSTSENNEg.PAPRGHLVRVEGNQRSEYMEDGNTLRHSVLVPYE 211
P53_HUMAN     160 AMAIYKQSQHMTEVVRRCPHHERCSD-SDg.LAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYE 222
P53_XENLA     134 ATAVYKKSEHVAEVVKRCPHHERSVEPGEd.AAPPSHLMRVEGNLQAYYMEDVNSGRHSVCVPYE 197
P53_ORYLA     145 ATAVYKKTEHVADVVRRCPHHQNEDS---..VEHRSHLIRVEGSQLAQYFEDPYTKRQSVTVPYE 204
Q27937_LOLFO  186 AMPIYMKPEHVQEVVKRCPNHATAKEHNEk.HPAPLHIVRCEHK-LAKYHEDKYSGRQSVLIPHE 248
P73_HUMAN     178 AMPVYKKAEHVTDVVKRCPNHELGRDFNEgqSAPASHLIRVEGNNLSQYVDDPVTGRQSVVVPYE 242
position          .   70    .   80    .   90    .  100    .  110    .  120    .
```

Due to the inherent variability of biological systems, a protein, such as the tumor suppressor p53 shown above, does not correspond to a specific sequence, but instead should be defined probabilistically as an evolutionary ensemble of sequences sharing a common cellular function. This ensemble can be modeled as a high dimensional probability distribution over every possible sequence. A generative statistical model, such as a **hidden Markov model (HMM)**, can 'emit' sequences with probabilities defined by such a distribution. HMMs are described in chapter 3.

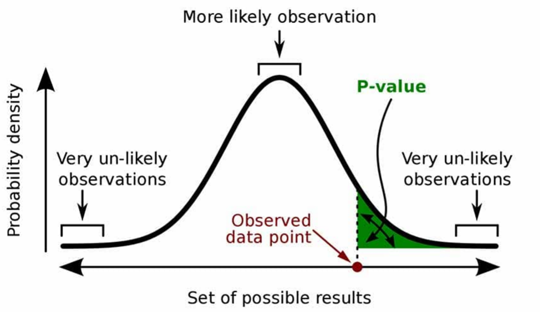


Representation of a protein probability distribution. The *x,y*-plane corresponds to a projection of the 'space' of all proteins onto 2-dimensions with similar sequences corresponding to points in the plane that are nearer to each other. The *z*-dimension plots the probability that each sequence (i.e., each point in the plane) is a specific type of protein, such as p53. The point of highest probability may be viewed as a consensus sequence for that protein.

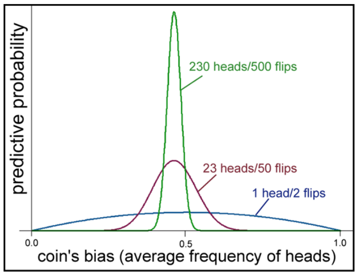
Similarity with classical genetics. Classical geneticists obtained statistical evidence for pairs of linearly ordered genes based on patterns of inherited traits in the absence of any direct cytological or molecular data. Likewise, one may obtain evidence for underlying protein properties based on patterns within sequence and structural data. Just as correlations among inherited traits may be due to the chromosomal locations of the associated genes, correlations among sequences and structures may be due to constraints imposed by a protein's function. And just as hypotheses generated by early genetic analyses have been validated by cytological and genomic studies, plausible hypotheses regarding protein function may be validated by biochemical and biophysical studies. However, just as a genetic analysis works well for some organisms (e.g., peas), but not for others (e.g., polyploid species such as hawkweed), our approach works better for some proteins than for others. Therefore, it is helpful to focus on those proteins most amenable to analysis—such as those belonging to large, taxonomically diverse superfamilies.

Frequentist vs Bayesian statistics. Here we utilize both frequentist and Bayesian statistics. A frequentist assigns probabilities to data, not to hypotheses. Hence, the frequentist looks for the probability of observing the data given a model (termed the null hypothesis). A frequentist often focuses on *p*-values. A *p*-value (shaded area in the figure on the right) is the probability of an observed or more extreme result assuming that the null hypothesis is true.

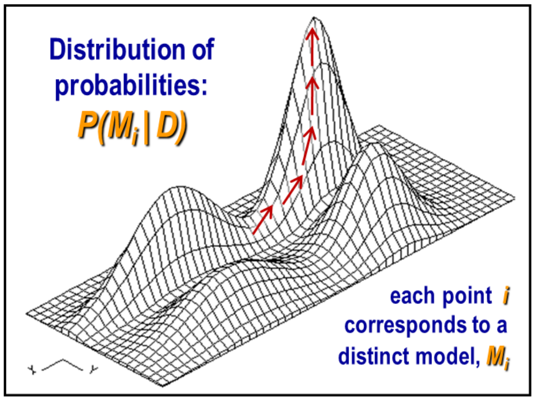
In contrast, Bayesian statistics assigns probabilities to hypotheses, arguing that all that we know for certain is the observed data; we cannot observe the underlying model generating the data. Bayesian statistics incorporates prior probabilities and updates model probabilities as more data become available. Prior probabilities provide a way to incorporate either general knowledge about the model or else no information at all, the latter being termed a non-informed prior where each outcome is treated as equally likely *a priori*.



Note that a Bayesian approach tells us what can reasonably be inferred given the available data. For example, consider a simple experiment that probabilistically predicts a coin's true property based on the observed data (i.e., the outcomes of flipping the coin). Such a probabilistic (or 'blurry') description most accurately represents what the data is telling us. Without such statistical criteria, one can easily be misled either by being too cautious or by not being cautious enough. The figure on the right shows the posterior probability distributions (over the possible biases of the coin) given three different observed data sets and given a non-informed prior.

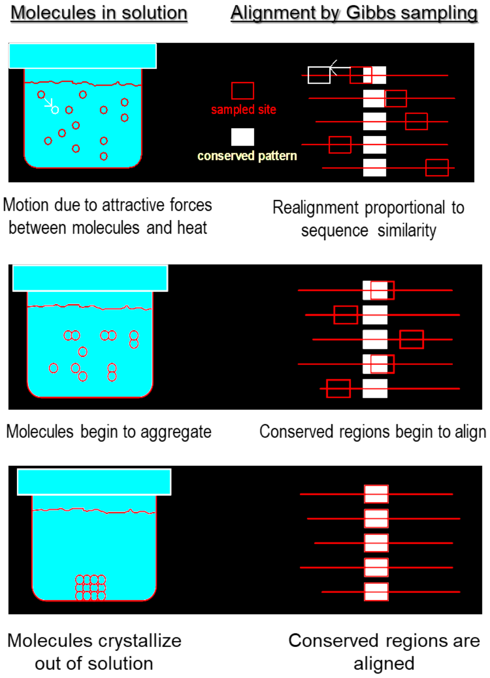


A Bayesian version of the scientific method. Our tools are largely based on Bayesian **Markov chain Monte Carlo (MCMC)** sampling, which may be viewed as a version of the scientific method that considers multiple alternative models concurrently: Instead of a single hypothesis regarding a specific model (termed M_i) being consistent or inconsistent with the outcome of a particular experiment, every possible model is assigned an implicit probability of being correct given the empirical observations (termed D), which in our case corresponds to sequence and structural data. Typically, one cannot compute probabilities for every model, so Bayesian procedures rely—as does the scientific method itself—on iterating between hypothesis testing and model refinement during MCMC sampling. In each iterative step, n_x alternative hypotheses regarding a particular modeled property, x, are evaluated (while other properties are held constant) by computing model probabilities over the full range of possible hypotheses for that property (i.e., $P(M_{x,j} \mid D)$ for $0 < j \leq n_x$). That property of the model is then updated according to the outcome of these 'experiments' (i.e., proportional to the computed probabilities for each hypothesis). MCMC sampling iterates through all model properties in this way until convergence (ideally) on the most probable models. When the underlying probability distribution is multimodal, as in the figure shown, the sampler can be initiated from multiple starting points to help find a global optimum. The optimum corresponds to a model that best describes (i.e., is most likely to have generated) the observed data. We use MCMC sampling in Chapters 3 to multiply align sequences and in Chapter 5 to hierarchically classify aligned sequences into subgroups, which we term Bayesian Partitioning with Pattern Selection (BPPS).


Distribution of probabilities:
$P(M_i \mid D)$
each point i corresponds to a distinct model, M_i
x
y

3. Multiple Sequence Alignment (MSA) via MCMC sampling

Finding conserved blocks via MCMC sampling. To illustrate MCMC sampling, the following describes how a program, termed the **Gibbs sampler**, identifies and aligns ungapped conserved segments, termed blocks, within a set of protein sequences. For simplicity, these blocks are of fixed length, The starting alignment consists of arbitrarily-selected candidate sites, one in each sequence, that will, of course, be very unlikely to correspond to a region most conserved across all sequences. Next, each site is realigned proportional to the probability that that site could have been generated by a model based on the other aligned sites; iteratively applying this operation favors the alignment of conserved over non-conserved regions and will ideally converge on an alignment of the most conserved region among all the sequences. This typically requires both shift operations to recenter the aligned blocks over the conserved regions and application of simulated annealing, where the sampling "temperature" is gradually lowered—that is, where sampling is distorted to favor higher probability sites more strongly over lower probability sites.


Molecules in solution
Alignment by Gibbs sampling
sampled site
conserved pattern
Motion due to attractive forces between molecules and heat
Realignment proportional to sequence similarity
Molecules begin to aggregate
Conserved regions begin to align
Molecules crystallize out of solution
Conserved regions are aligned

The process of sequence alignment via Gibbs sampling is analogous to molecular crystallization as explained by statistical thermodynamics (see figure to the right). Starting from an arbitrary 'state' a thermodynamic 'system' evolves due to internal 'forces' and random fluctuations where the attractive forces gradually bring similar atoms or molecules together. Hence, just as some molecules begin to aggregate during crystallization, some conserved regions begin to align during Gibbs sampling. Given an appropriate temperature, eventually the 'system' reaches equilibrium where, just as molecules crystalize out of solution, the conserved regions align. It is called Gibbs sampling in reference to this analogy between the sampling algorithm and statistical physics.

Fig. 3.1 shows a Gibbs sampler generated alignment consisting of 30 helix-turn-helix (HTH) DNA-binding sites present within otherwise very different proteins, any two of which lack significant pairwise sequence similarity. The sampler easily finds these HTH sites with high statistical significance.

```
Sigma-37         223 IIDLTYIQNK SQKETGDILGISQMHVSR LQRKAVKKLR  240    A25944
SpoIIIC           94 RFGLDLKKEK TQREIAKELGISRSYVSR IEKRALMKMF  111    A28627
NahR              22 VVFNQLLVDR RVSITAENLGLTQPAVSN ALKRLRTSLQ   39    A32837
Antennapedia     326 FHFNRYLTRR RRIEIAHALCLTERQIKI WFQNRRMKWK  343    A23450
NtrC (Brady.)    449 LTAALAATRG NQIRAADLLGLNRNTLRK KIRDLDIQVY  466    B26499
DicA              22 IRYRRKNLKH TQRSLAKALKISHVSVSQ WERGDSEPTG   39    B24328 (BVECDA)
MerD               5       MNAY TVSRLALDAGVSVHIVRD YLLRGLLRPV   22    C29010
Fis               73 LDMVMQYTRG NQTRAALMMGINRGTLRK KLKKYGMN     90    A32142 (DNECFS)
MAT a1            99 FRRKQSLNSK EKEEVAKKCGITPLQVRV WFINKRMRSK  116    A90983 (JEBY1)
Lambda cII        25 SALLNKIAML GTEKTAEAVGVDKSQISR WKRDWIPKFS   42    A03579 (QCBP2L)
Crp (CAP)        169 THPDGMQIKI TRQEIGQIVGCSRETVGR ILKMLEDQNL  186    A03553 (QRECC)
Lambda Cro        15 ITLKDYAMRF GQTKTAKDLGVYQSAINK AIHAGRKIFL   32    A03577 (RCBPL)
P22 Cro           12 YKKDVIDHFG TQRAVAKALGISDAAVSQ WKEVIPEKDA   29    A25867 (RGBP22)
AraC             196 ISDHLADSNF DIASVAQHVCLSPSRLSH LFRQQLGISV  213    A03554 (RGECA)
Fnr              196 FSPREFRLTM TRGDIGNYLGLTVETISR LLGRFQKSGM  213    A03552 (RGECF)
HtpR             252 ARWLDEDNKS TLQELADRYGVSAERVRQ LEKNAMKKLR  269    A00700 (RGECH)
NtrC (K.a.)      444 LTTALRHTQG HKQEAARLLGWGRNTLTR KLKELGME    461    A03564 (RGKBCP)
CytR              11 MKAKKQETAA TMKDVALKAKVSTATVSR ALMNPDKVSQ   28    A24963 (RPECCT)
DeoR              23 LQELKRSDKL HLKDAAALLGVSEMTIRR DLNNHSAPVV   40    A24076 (RPECDO)
GalR               3         MA TIKDVARLAGVSVATVSR VINNSPKASE   20    A03559 (RPECG)
LacI               5       MKPV TLYDVAEYAGVSYQTVSR VVNQASHVSA   22    A03558 (RPECL)
TetR              26 LLNEVGIEGL TTRKLAQKLGVEQPTLYW HVKNKRALLD   43    A03576 (RPECTN)
TrpR              67 IVEELLRGEM SQRELKNELGAGIATITR GSNSLKAAPV   84    A03568 (RPECW)
NifA             495 LIAALEKAGW VQAKAARLLGMTPRQVAY RIQIMDITMP  512    S02513
SpoIIG           205 RFGLVGEEEK TQKDVADMMGISQSYISR LEKRIIKRLR  222    S07337
Pin              160 QAGRLIAAGT PRQKVAIIYDVGVSTLYK TFPAGDK     177    S07958
PurR               3         MA TIKDVAKRANVSTTTVSH VINKTRFVAE   20    S08477
EbgR               3         MA TLKDIAIEAGVSLATVSR VLNDDPTLNV   20    S09205
LexA              27 DHISQTGMPP TRAEIAQRLGFRSPNAAE EHLKALARKG   44    S11945
P22 cI            25 SSILNRIAIR GQRKVADALGINESQISR WKGDFIPKMG   42    B25867 (Z1BPC2)
                                  ***************** ***
```

Fig. 3.1. Gibbs sampler alignment of helix-turn-helix DNA-binding sites common to 30 proteins. Columns from left to right are: sequence name; residue positions of the left end of each sequence; left flanking region; 18-residue conserved blocks; right flanking region; residue positions of the right-ends of each sequence; NBRF/PIR accession number; and NBRF/PIR code name, if available. Asterisks (***) below the alignment indicate a 20-residue segment based on structural superpositions. Almost equal values of information per parameter were given by block widths of 18 to 21 residues: the longer widths extended to the right the 18-residue block shown. Taken from reference {x?}.

Ungapped internal repeats within bacterial porins. In addition to identifying a fixed number of conserved patterns in each input sequence, the Gibbs sampler can also identify an undefined number of ungapped internal repeats in each sequence. This is illustrated in Fig. 3.2 for bacterial porins, for which the sampler found a conserved pattern associated with alternate β-strands contacting the lipid membrane within the porin homotrimeric complex (Fig. 3.2C). Detection of these strands is based on an additional feature of MCMC sampling, namely that it can provide a measure of model uncertainty (as was illustrated by the flipped coin example above). Specifically, this involves continuing to sample among alternative models after convergence, in which case uncertain aligned regions will tend to be sampled out of the alignment more often whereas more certain aligned regions will tend to be retained. The frequency with which each segment is sampled estimates the predictive probability of that segment belonging to the associated statistical model (column 3 in **Fig. 3.2A**). To avoid converging on a highly conserved region shared by a few closely related sequences, it is important to remove close homologs prior to an analysis, which can be done using the *purge* program.

A

segment	site	prob.	protein
DAQEMAVAAAYTF	146	0.9605	PORI_RHOCA
DMEQLELAAIAKF	180	0.8575	
DVTYYGLGASYDL	259	0.9960	
SDMVADLGVKFKF	289	0.5690	
KAEQWATGLKYDA	232	0.9650	OMPF_ECOLI
KTQDVLLVAQYQF	275	0.9955	
LVNYFEVGATYYF	313	0.9910	
SDDTVAVGIVYQF	350	0.9980	
DNDIAFVGAAYKF	190	0.9980	PORI_RHOBL
AGDQVTLYGNYAF	220	0.7770	
ADTAYGIGADYQF	249	0.9635	
NETVADVGVRFDF	277	0.9650	

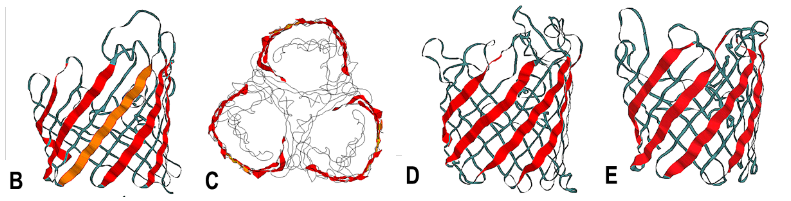
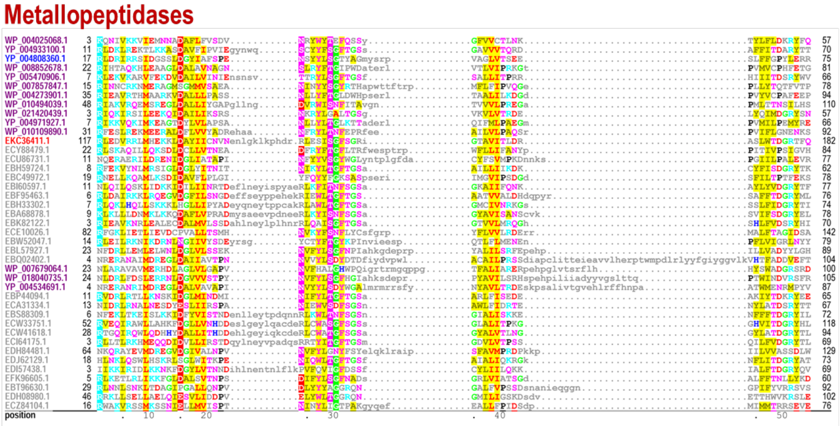
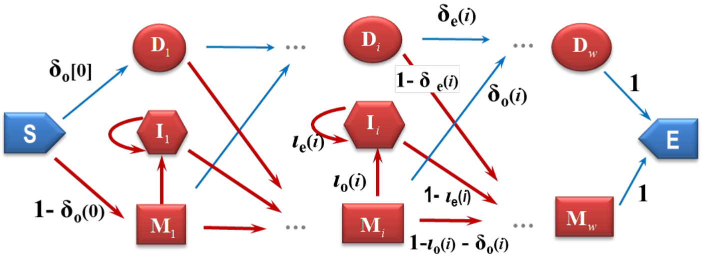


Fig. 3.2. Conserved internal repeats detected in bacterial porins. **A**. The alignment of repeats in three porins of known structure: PORI-RHOCA (*R. capsulatus* porin), OMPF-ECOLI (*E. coli* OmpF), and PORI-RHOCA (*R. blastica* porin). These repeats along with repeats in sixteen other distantly related bacterial porins were detected by the Gibbs sampler. The predictive probability that each aligned segment contains the repeat is given in column 3 and is based on the fraction of times that segment was sampled after convergence. **B-E:** Locations of conserved repeats within bacterial porins. Tracing of α-carbons are shown as ribbons or strands. Highlighted in red are those β-strands in bacterial porins that the Gibbs sampler assigned predictive probabilities of ≥ 0.50. These porins form homotrimers, arranged as shown in panel C, with conserved repeats corresponding to non-interacting β-strands. **B:** *R. capsulatus* porin (3POR). The segment at site 228 in PORI-RHOCA ("DHKAYGLSVDSTF") is highlighted in orange because of its relatively high predictive probability (p = 0.258) (by default only repeats with $p \geq 0.50$ are reported). **C:** Homotrimeric *R. capsulatus* porin complex showing that the conserved repeats occur at the membrane interface. **D:** *E. coli* OmpF porin (IOMF). **E:** *R. blastica* porin (IPRN).

Gibbs Sampler for Multi-alignment Optimization (GISMO). GISMO uses MCMC sampling to search for a protein **hidden Markov Model** (**HMM**) that is most likely to have generated a set of input sequences [9,10]. In the process, it multiply aligns the sequences while also inferring position specific gap penalties. As the figure on the right illustrates, it tends to align sequences as multiple conserved regions and can easily span across large insert regions without fragmenting the rest of the alignment. Unlike most MSA programs, GISMO will not align unrelated (e.g., randomly generated) sequences.



An HMM is an abstract statistical model that probabilistically generates strings of characters in an alphabet that for protein sequences consists of the 20 amino acid residues. The architecture of a GISMO HMM is shown on the right. It consists of 5 types of states: a start (S) and an end (E) state, and delete (D), insert (I) and match (M) states; and of transition arrows between states. Transition arrows have assigned probabilities such that the sum of the probabilities out of each state sum to 1. An HMM generates a sequence by stochastically following a path through the HMM: beginning in the S state, subsequent states are sampled based on the transition probabilities; upon transition into an insertion (I) or a match (M) state, the HMM emits a character in the alphabet (i.e., an amino acid residue) based on the emission probabilities assigned to that state. This process is terminated upon reaching the end state (E) with the output being the sequence of residues emitted.



There are three algorithmic operations associated with such an HMM: (1) Scoring an arbitrary sequence against an HMM as the probability of the HMM having generated that sequence. This requires summing probabilities over all possible paths through the HMM that would have generated that sequence; this is done using the **forward algorithm**, which, like the Smith-Waterman (SW) algorithm [11], is based on dynamic programing. (2) Optimally aligning a given sequence against an HMM by finding the path through the HMM that is most likely to have generated that sequence. This is done using the Viterbi algorithm, which is also based on dynamic programing. (3) Finding the HMM most likely to have generated a given set of input sequences—in other words, to derive the architecture of this HMM and the maximum likelihood estimate of its transition and emission probabilities. Although no tractable exact algorithm is known for this, one can use Markov chain Monte Carlo sampling to obtain a near optimal solution.

GISMO uses MCMC sampling in this way to search for an optimal HMM and an associated optimal MSA by probabilistically sampling HMM architectures and corresponding parameter settings based on a given set of input sequences. This process begins by arbitrarily aligning the sequences and then using this alignment to parameterize an initial HMM, termed M_0. An HMM is parameterized by estimating transition and emission probabilities based on the observed residue and indel frequencies in the MSA, where each aligned column contains those residues emitted upon transitions into a corresponding match (M) state and gap ('-') 'residues' for transitions into a delete (D) state, respectively, and where each insertion consists of those residues emitted upon transition into a corresponding insert (I) state. Next, GISMO iteratively samples each sequence's alignment within the MSA based on an HMM derived from the MSA without that sequence. It then re-parameterizes the evolving HMM based on the sampled MSA. Specifically, in the i^{th} iteration GISMO computes the probability of each of n_x alternative (slightly different) HMMs having generated the input sequence data D (i.e., it computes $P(M_{i,j} \mid D)$ for $0 < j \leq n_x$). It then samples one $M_{i,j}$ (and a corresponding MSA) proportional to these probabilities—where each $M_{i,j}$ and MSA corresponds to a distinct realignment of the sampled sequence. Other MCMC operations sample aligned columns in and out of the MSA. Finally, GISMO stops upon convergence on a high probability (ideally, optimal) HMM and corresponding MSA.

On average, GISMO aligned 408 benchmark sequence sets more accurately than did six of the most popular MSA programs (panel A in **Fig. 3.3**)—especially for the most difficult to align sequence sets. Unlike these other programs, which are deterministic and thus return the same MSA for each run, GISMO is stochastic (panel B). However, the program-to-program variability (panel C) is greater than the run-to-run variation for GISMO (panel B).

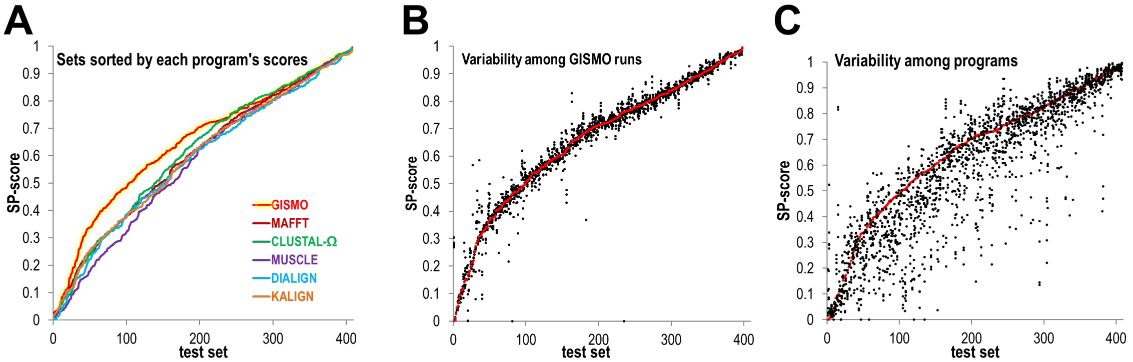
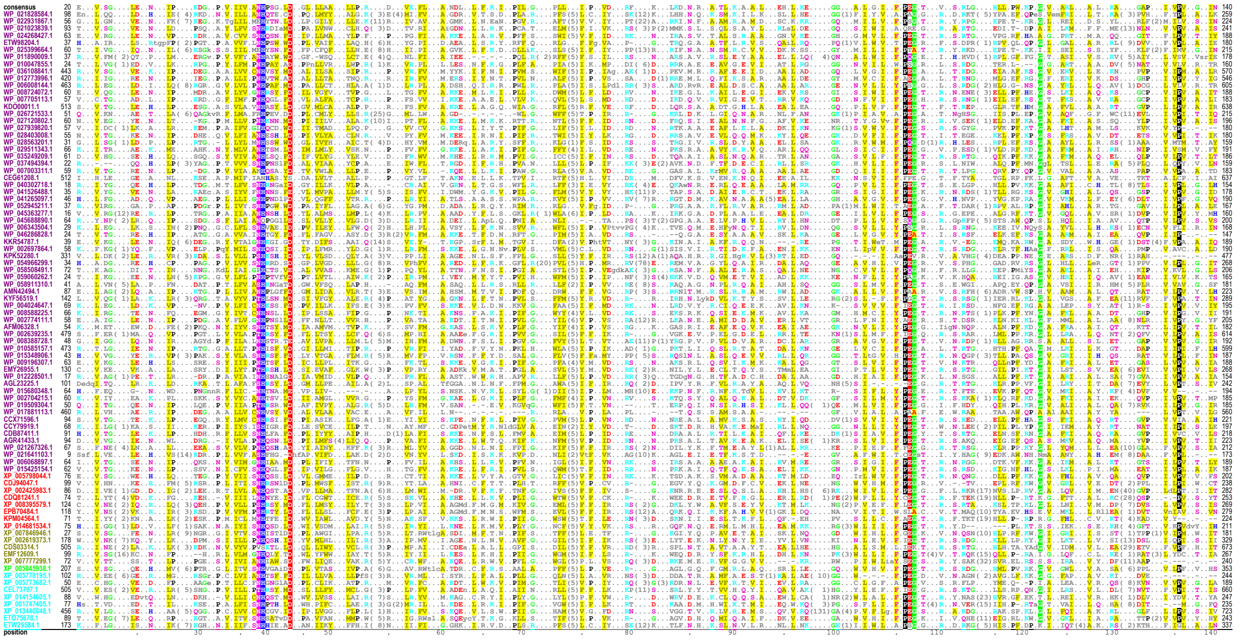
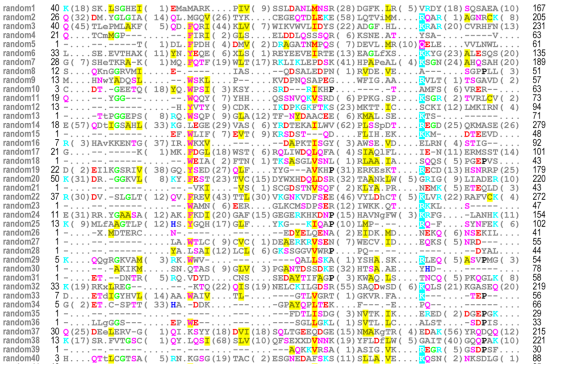


Figure 3.3. Alignment quality among various MSA programs based on sum of pair (SP-)scores, which vary from 0 (no correctly aligned sequence pairs) to 1 (all pairs aligned correctly). **A**. The sorted (lowest to highest) SP-scores obtained by six MSA programs. **B**. Run-to-run variability in SP-scores over six GISMO runs. Test set data points are sorted along the x-axis by the SP-score obtained for each set on the first run (red data points) out of six. **C**. SP-scores for these six programs sorted by the GISMO score on each test set. GISMO SP-scores (for a single run) are shown in red. Each red data point and the five black data points (one point for each program) plotted in the same column correspond to the same test sequence set.

GISMO can run a user-specified number of CPU threads in parallel, which allows for a more thorough exploration of the posterior probability distribution over possible HMMs. Rather than speeding up the program, this is designed to improve MSA accuracy by improving its ability to find a more nearly optimal MSA. **Appendix 1** provides mathematical details regarding the GISMO statistical model and algorithm. The **GAMBIT** program will attempt to further improve a GISMO-generated MSA by continuing to apply MCMC sampling, which can also provide a measure of alignment uncertainty (as was illustrated above in Fig. 2.2A). As described below, the **eCOMPASS** (evaluative Comparison of Multiple Protein Alignments by Statistical Score) program [12] evaluates the relative quality of two alignments of the same sequences based on direct coupling analysis (DCA) as an alternative to using Sum-of-Pairs (SP) scores.

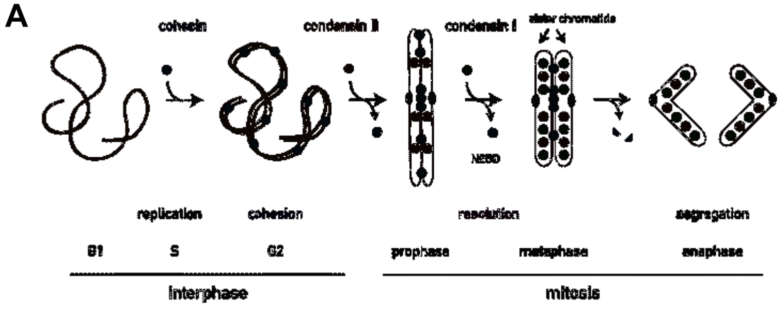
A biomedical application. An early MCMC sampler implemented in conjunction with a database search procedure identified very subtly conserved motifs within sequences that included lysophospholipid acyltransferases. Among the sequences found was tafazzin {PMID: 9259571}, which is associated with Barth syndrome, an inherited cardiomypathic disorder in children {PMID: 6142097; PMID: 8739954}. This suggested that Barth syndrome was due to an acyltransferase deficiency, which was later clinically confirmed (see J Pediatr. 2002. 141(5): 729-733); this led to potential treatments for this disease (see J Lipid Res. 2003. 44(3): 560-566), including the first-ever FDA approved treatment: Elamipretide, a small mitochondrially-targeted tetrapeptide that appears to reduce the production of toxic reactive oxygen species, thereby stabilizing cardiolipin. Abnormal cardiolipin is a specific diagnostic marker of cardiomyopathies caused by Barth syndrome mutations leading to alterations in the fatty acid composition of several phospholipids. The latest version of GISMO creates the following, improved alignment of these acyltransferase proteins:

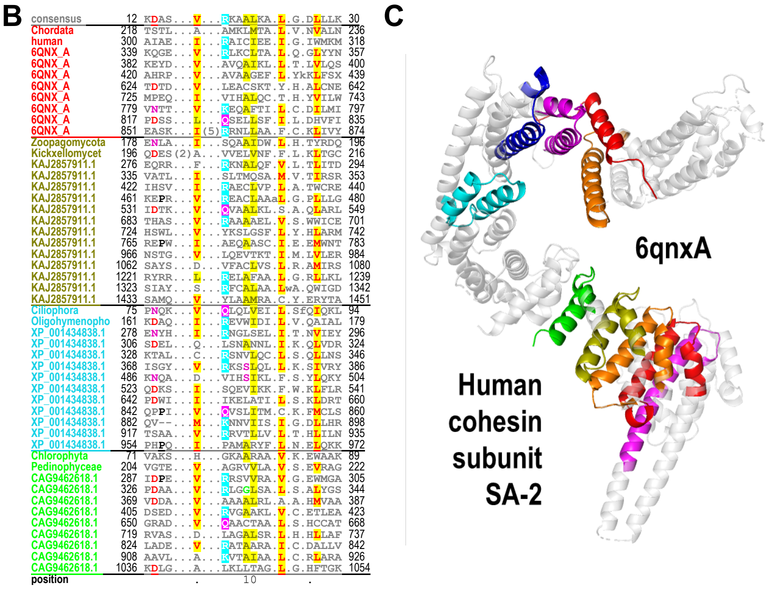




Statistical significance. A key feature of GISMO, which most MSA programs lack, is that it will not align sequences in the absence of statistically significant sequence similarity—though it will align subtly conserved regions that might appear to lack significant similarity. Shown on the right, for example, is a MAFFT alignment of randomly shuffled sequences, which could be misinterpreted as biologically relevant, and which GISMO does not align. To best distinguish subtle sequence similarities from noise, GISMO should be applied to input sets with closely related sequences removed using the **purge** option within our **tweakseq** program.

Protein Short Internal Repeat Detection (PSIRD). The PSIRD program detects subtly conserved *gapped* repeats by first applying the Gibbs sampler to detect and output ungapped repeats along with flanking regions. This output is then used as input to GISMO to create a gapped alignment of the repeat regions. This alignment is then used as input to a gapped search procedure that will both detect and output additional (gapped) repeats. (The user may apply an option that requires a minimal number of internal repeats in each sequence.) This output is then used as input to GISMO once again to align the enlarged repeat set. Several rounds of this strategy may be applied to expand the final alignment, which may then include distant members of the repeat superfamily—often including previously unknown members.



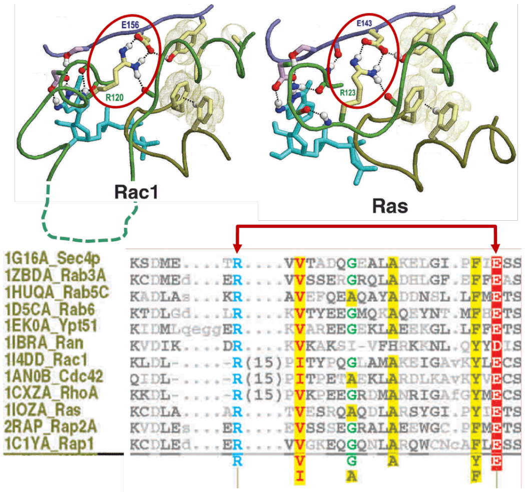


```
consensus           12 KDAS...V...RKAALKA.L.G.DLLLK 30
Chordata           218 TSTL...A...AMKLMTA.L.V.NVALN 236
human              300 AIAE...I...RAICIEE.I.G.IWMKM 318
6QNX_A             339 KQGE...V...RLKCLTA.L.Q.GLYYN 357
6QNX_A             382 KEYD...V...AVQAIKL.L.T.LVLQS 400
6QNX_A             420 AHRP...V...AVAAGEF.L.YkKLFSX 439
6QNX_A             624 TDTD...V...LEACSKT.Y.H.ALCNE 642
6QNX_A             725 MPEQ...I...VIHALQC.T.H.YVILW 743
6QNX_A             779 VNTT...V...KEQAFTI.L.C.DILMI 797
6QNX_A             817 PDSS...L...QSELLSF.I.L.DHVFI 835
6QNX_A             851 EASK...I(5)RRNLLAA.F.C.KLIVY 874
Zoopagomycota      178 ENLA...I...SQAAIDW.L.H.TYRDQ 196
Kickxellomycet     196 QDES(2)A...VVELVNF.F.L.KLTGC 216
KAJ2857911.1       276 EQRR...F...RKNALQF.V.L.TLITD 294
KAJ2857911.1       335 VATL...I...SLTMQSA.M.V.TIRSR 353
KAJ2857911.1       422 IHSV...I...RAECLVP.L.A.TWCRE 440
KAJ2857911.1       461 KEPR...V...REACLAAaL.G.PLLLG 480
KAJ2857911.1       531 IDTK...V...QVAALKL.S.A.QLARL 549
KAJ2857911.1       683 THAS...V...RAAAAEL.V.S.WWICE 701
KAJ2857911.1       724 HSWL...V...YKSLGSF.L.Y.HLARM 742
KAJ2857911.1       765 REPW...I...AEQAASC.I.E.EMWNT 783
KAJ2857911.1       966 NSTG...V...LQEVTKT.I.M.LVLER 984
KAJ2857911.1      1062 SAYS...D...VFACLVS.L.R.AMIRS 1080
KAJ2857911.1      1221 RYRR...L...RELAFAA.L.G.RLLKL 1239
KAJ2857911.1      1323 SIAY...S...RFCALAA.LwA.QWIGD 1342
KAJ2857911.1      1433 SAMQ...V...YLAAMRA.C.Y.ERYTA 1451
Ciliophora          75 PNQK...V...QLQLVEI.L.SfQIQKL 94
Oligohymenopho     161 KDAQ...I...REVWIDI.L.V.QAIAL 179
XP_001434838.1     278 ENYH...I...RNGLSEL.I.T.NVIEY 296
XP_001434838.1     306 SDEL...Q...LSNANNL.I.K.QLVDR 324
XP_001434838.1     328 KTAL...C...RSNVLQC.L.S.QLLNS 346
XP_001434838.1     368 ISGY...V...RKSSLQL.L.K.SIVRY 386
XP_001434838.1     486 KNQA...D...VIHSIKL.F.S.YLQKY 504
XP_001434838.1     523 QDKS...I...SQEVIKK.F.W.KLFLR 541
XP_001434838.1     642 PDWI...I...IKELATI.L.S.KLDRT 660
XP_001434838.1     842 QPPI...V...QVSLITM.C.K.FMCLS 860
XP_001434838.1     882 QV--...M...KNNVIIS.I.G.DLLHR 898
XP_001434838.1     917 TSAA...V...RRVTLLV.L.T.HLILN 935
XP_001434838.1     954 PHPQ...I...PAMARYF.L.N.ELQKK 972
Chlorophyta         71 VAKS...H...GKAARAA.V.K.EWAAK 89
Pedinophyceae      204 VGTE...V...AGRVVLA.V.S.EVRAG 222
CAG9462618.1       287 IDPE...V...RRSVVRA.V.G.EWMGA 305
CAG9462618.1       326 PDAA...V...RLGGLSA.L.S.ALYGS 344
CAG9462618.1       369 VDDA...V...AAAALRL.A.A.HMVAA 387
CAG9462618.1       405 DSED...V...RVGAAKL.V.C.ETLEA 423
CAG9462618.1       650 GRAD...V...QAACTAA.L.S.HCCAT 668
CAG9462618.1       719 RVAS...D...LAGALSR.L.H.HLLAF 737
CAG9462618.1       824 LADE...V...RATAARA.I.C.DALLV 842
CAG9462618.1       908 AAVL...A...KVTAIAA.L.C.RLARA 926
CAG9462618.1      1036 KDLG...A...LKLLTAG.L.G.HFTGK 1054
position                      .      10      .
```

Fig. 2.3. Condensin and cohesion HEAT repeats. **A**. Roles of condensin and cohesion in the cell cycle. **B**. HEAT-like repeats in subunits of these complexes. See Fig. 1.2 for color scheme. **C**. The structural locations of the repeats detected within human cohesinSA-2 (pdbid: 6qnxA).

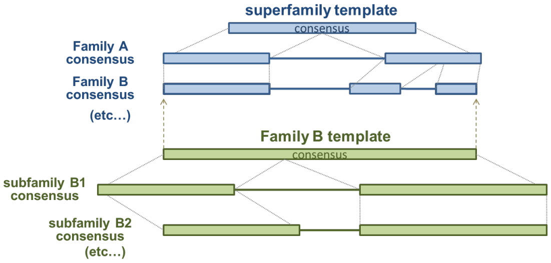
This is illustrated by our detection of HEAT repeats in chromosome-associated protein components [4] (**Fig. 2.3**), which led to our discovery of a new vertebrate condensin based on the presence of similar HEAT repeats in one of its subunits [6]. Note, however, that to detect subtly conserved regions in a set of protein sequences, it is again important to eliminate closely related sequences, which would otherwise introduce strong signals to mislead the sampler. This is done either using the **purge** [7] or the (heuristic) **cd-hit** procedure (by Weizhong Li) [8] both of which are implemented within our **tweakseq** program. Likewise, **Fig. 2.4** shows a PSIRD analysis of β-propeller repeats in UV-damage DNA-binding (UVDDB) proteins. An very early version of PSIRD predicted the presence of these UVDDB repeats [3], which was completely unknown at the time but was subsequently confirmed (six years later) through crystal structure analysis [5]. Although the prediction of structural repeats is no longer difficult with the advent of AlphaFold structure predictions, characterizing such repeats is important to our goal of decomposing sequence constraints with a view to deciphering protein functional determinants.

4. Sequence detection & alignment using a hierarchical MSA


E156
R120
Rac1
E143
R123
Ras
1G16A_Sec4p KSDME....TR....VVTADQGEALAKELGI.PFIESS
1ZBDA_Rab3A KCDMEd...ER....VVSSERGRQLADHLGF.EFFEAS
1HUQA_Rab5C KADLAs...KR....AVEFQEAQAYADDNSL.LFMETS
1D5CA_Rab6 KTDLGd...LR....KVTYEEGMQKAQEYNT.MFHETS
1EK0A_Ypt51 KIDMLqeggER....KVAREEGEKLAEEKGL.LFFETS
1IBRA_Ran KVDIK....DR....KVKAKSI-VFHRKKNL.QYYDIS
1I4DD_Rac1 KLDL-....-R(15)PITYPQGLAMAKEIGAvKYLECS
1AN0B_Cdc42 QIDL-....-R(15)PITPETAEKLARDLKAvKYVECS
1CXZA_RhoA KKDL-....-R(15)PVKPEEGRDMANRIGAfGYMECS
1IOZA_Ras KCDLA....AR....TVESRQAQDLARSYGI.PYIETS
2RAP_Rap2A KVDLEs...ER....EVSSSEGRALAEEWGC.PFMETS
1C1YA_Rap1 KCDLEd...ER....VVGKEQGQNLARQWCNcAFLESS
R V G A Y E
I A F

Even the best MSA programs will fail to correctly align some protein sequences. Various Ras-like GTPases, for example, conserve an Arg-Glu salt bridge where the arginine is sometimes preceded by a deletion or an insertion and is sometimes followed by a 15-residue insertion. Hence, aligning this arginine correctly is impossible without manual curation, which is prohibitively time consuming for all but the smallest MSAs.

MAPGAPS. The **MAPGAPS (Multiply-Aligned Profiles for Global Alignment of Protein Sequences)** program [14] addresses this problem by detecting and aligning database sequences using as the query a curated hierarchical MSA (hiMSA). An shown schematically in chapter 1, a hiMSA consists of a set of MSAs, one for each subgroup within a given superfamily, that are multiply aligned to each other and to the superfamily within a 'template' MSA, as shown schematically on the left. MAPGAPS first converts the hiMSA into a set of multiply aligned profiles, which are then used both to detect and globally align database sequences related to the superfamily. It relies on Karlin–Altschul statistics as a measure of significance and on PSI-BLAST (and other) heuristics for speed. The BLAST and PSI-BLAST algorithms are described as follows:



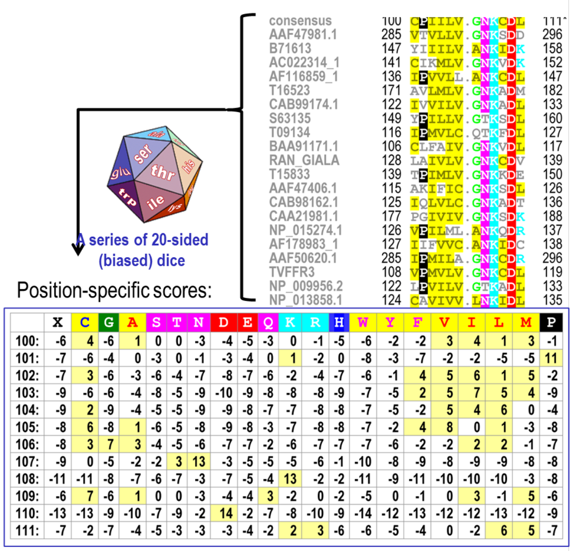
BLAST (basic local alignment search tool) applies a heuristic algorithm to rapidly identify database sequences with statistically significant pairwise similarity to a query sequence. This involves the following steps: 1. Create a lookup table for each 3-letter "word" in the query sequence. 2.Using the lookup table determine the pairwise score between each 3-letter word in the query and each 3-letter word in each database sequence. 3.Extend each word-to-word pair with a score above a specified threshold into an ungapped pairwise alignment (termed a segment). 4.For each high-scoring pair (HSP) (i.e., segments with scores greater than a second specified threshold) perform a (gapped) Smith-Waterman pairwise alignment. 5.Return SW-alignments with statistically significant scores.

As a measure of significance, BLAST relies on Karlin-Altschul (extreme value) statistics to calculate the probability by chance of obtaining a given pairwise score. By default, such scores are calculated using a BLOSUM62 substitution matrix, which represents a log-odds score for residue *i* substituting for residue *j* within a homologous protein; this is based on the relative substitution frequencies (i.e., probabilities) within aligned (ungapped) conserved regions (termed 'blocks') in related proteins. BLOcks SUbstitution Matrices (BLOSUM) are a family of substitution matrices used in bioinformatics to assess the similarity between amino acids in protein sequences, with the "62" prefix indicating the percentage identity of the sequences used to create that matrix. For example, the BLOSUM62 score between an Asp and Glu residue (denoted as $S_{D,E}$) is positive because an (acidic) aspartate residue is more likely (than is expected by chance) to mutate into an (acidic) glutamate residue and, as a result, this transition is less disruptive biochemically. Specifically, this score is calculated as:

$$S_{D,E} = \log_2 \frac{p_D \cdot P_{D,E}}{p_D \cdot p_E} = \log_2 \frac{P_{D,E}}{p_E} = \log_2 \frac{0.0868}{0.051} = \log_2 1.69 = 0.755$$

where p_D and p_E represent the overall probability of observing Asp or Glu, respectively, and $P_{D,E}$ represents the probability of Asp substituting for Glu or vice versa. The BLOSUM62 score between Phe and Glu is negative because an (aromatic) phenylalanine residue is less likely (than is expected by chance) to mutate into glutamate because this transition is likely to be biochemically disruptive.

PSI-BLAST. A BLAST search creates an implicit MSA where all the detected database sequences are aligned against the query sequence, which serves as a template. PSI-BLAST converts this alignment into a **position-specific scoring matrix (PSSM)**, where the *i*th row of the scoring matrix gives the log-likelihood ratio that, at the *i*th position of the alignment, an aligned protein sequence harbors the amino acid given in the *j*th column of the matrix. Each likelihood ratio is calculated by dividing the observed frequency of that residue in the *i*th column of the alignment by the overall frequency of that residue in proteins as a whole. This is illustrated in the figure on the right. Using a PSSM obtained from an initial BLAST search in this way, PSI-BLAST performs a sequence-to-profile search using this profile as the query for a second BLAST search which typically will detect additional sequences due to the improved protein sequence model. From this search yet another PSSM is obtained and the same process is repeated either for a user specified number of iterations or until convergence (i.e., until no additional sequences are detected). The PSI-BLAST program can save the PSSM, which may be used as a query to initiate a database search. **MAPGAPS** uses this feature to initiate a PSI-BLAST search using an hiMSA as the query.



```
consensus     100  CPIILV.GNKCDL  111
AAF47981.1    285  VTVLLV.GNKSDD  296
B71613        147  YIIILV.ANKIDK  158
AC022314_1    141  CIKMLV.GNKVDK  152
AF116859_1    136  IPVVLL.ANKCDL  147
T16523        171  AVLMLV.GNKADM  182
CAB99174.1    122  IVVILV.GNKADL  133
S63135        149  YPILLV.GTKSDL  160
T09134        116  IPMVLC.QTKFDL  127
BAA91171.1    106  CLFAIV.GNKVDL  117
RAN_GIALA     128  LAIVLV.GNKCDV  139
T15833        139  TPIMLV.GNKKDE  150
AAF47406.1    115  AKIFIC.GNKSDL  126
CAB98162.1    125  IQLVLC.GNKADT  136
CAA21981.1    177  PGIVIV.GNKSDK  188
NP_015274.1   126  VPILML.ANKQDR  137
AF178983_1    127  IIFVVC.ANKIDC  138
AAF50620.1    285  IPMILA.GNKCDR  296
TVFFR3        108  VPMVLV.GNKCDL  119
NP_009956.2   122  LPVILV.GTKADL  133
NP_013858.1   124  CAVIVV.LNKIDL  135
```

	X	C	G	A	S	T	N	D	E	Q	K	R	H	W	Y	F	V	I	L	M	P
100:	-6	4	-6	1	0	0	-3	-4	-5	-3	0	-1	-5	-6	-2	-2	3	4	1	3	-1
101:	-7	-6	-4	0	-3	0	-1	-3	-4	0	1	-2	0	-8	-3	-7	-2	-2	-5	-5	11
102:	-7	3	-6	-3	-6	-4	-7	-8	-7	-6	-2	-4	-7	-6	-1	4	5	6	1	5	-2
103:	-9	-6	-6	-4	-8	-5	-9	-10	-9	-8	-8	-8	-9	-7	-5	2	5	7	5	4	-9
104:	-9	2	-9	-4	-5	-5	-9	-9	-8	-7	-7	-8	-8	-7	-5	-2	5	4	6	0	-4
105:	-8	6	-8	1	-6	-5	-8	-9	-8	-3	-7	-8	-8	-7	-2	4	8	0	1	-3	-8
106:	-8	3	7	3	-4	-5	-6	-7	-7	-2	-6	-7	-7	-7	-6	-2	-2	2	2	-1	-7
107:	-9	0	-5	-2	-2	3	13	-3	-5	-5	-5	-6	-1	-10	-8	-9	-8	-9	-9	-8	-8
108:	-11	-11	-8	-7	-6	-7	-3	-7	-5	-4	13	-2	-2	-11	-9	-11	-10	-10	-10	-3	-8
109:	-6	7	-6	1	0	0	-3	-4	-4	3	-2	0	-2	-5	0	-1	0	3	-1	5	-6
110:	-13	-13	-9	-10	-7	-9	-2	14	-2	-7	-8	-10	-9	-14	-12	-13	-12	-12	-13	-12	-9
111:	-7	-2	-7	-4	-5	-3	-3	-3	-4	-2	2	3	-6	-6	-5	-4	0	-2	6	5	-7

Running MAPGAPS. The NCBI maintains a conserved domain database (CDD) [15] of manually curated hiMSAs for various protein superfamilies that can be downloaded as input for MAPGAPS [16]. These can be converted into MAPGAPS-readable format using the **cdd2mgs** program. We typically perform a MAPGAPS search on the NCBI fasta-formated non-redundant (nr) [17] and pdbaa databases. The **addphyla** program annotates these input sequences with taxids, and with class, phylum and kingdom designations—information that is used by our other programs. This requires downloading the NCBI taxdump.tar.gz and prot.acession2taxid.gz files [18] and, for the pdbaa sequences, the pdb.accession2taxid file. Taxonomic information is added after each sequence identifier (<seq_id>) using the syntax: `<seq_id> {|s(taxid)|p;c(k)>}` where ***s*** is the starting residue position for N-terminal truncated sequences; ***taxid*** is the sequence's NCBI taxonomy identifier as an unsigned positive integer; ***p*** and ***c*** are the names of the phylum and class, respectively; and $\boldsymbol{k} \in \{A, B, V, M, F, E, H\}$ indicates the kingdom, where *A*=archaea, *b*=bacteria, *V*=plants, *M*=metazoa, *F*=fungi, *E*=protozoa, and *H*=virus.

To speed up a MAPGAPS search and alignment, it is best to split up the database sequences into smaller sets (of ≤ 250,000 sequences) using our **twkseq** program with the 'split' command and then run a grid node MAPGAPS search on each set separately. MAPGAPS generated MSAs can be concatenated into a single file and then merged into one MSA using the **twkcma** 'merge' command. Note that our programs create and utilize MSAs in **cma-format**, which is more compact than other formats and thereby facilities the creation of very large MSAs.

The **convert_msa** program converts MSAs from cma to fasta format and vice versa or from cma to rich text format. Often cma-formatted MSAs require some additional processing prior to being used as input by our programs; for this **twkcma** can again be used. For example, it is recommended that sequence redundancy be reduced to between 90-95% sequence identity using the 'U' command. To remove redundant sequences faster use the cdhit' command (a heuristic method) or (using an exhaustive algorithm) the 'U' command (an exhaustive method) with the -thrds option for multithreading. The mincol=<real> command removes sequence fragments. The following table describes these and other commonly used **twkcma** options:

Command	Description
cdhit=<int>	heuristically remove all but one sequence among those sharing ≥ <int> percent identity using cd-hit (range: 40-100)
csq	output a cma formatted consensus sequence corresponding to the input MSA
hsw	create a *.hsw file that down weights aligned sequences for redundancy (as required by various programs)
merge	merge concatenated MSAs (all having the same number of aligned columns) into a single MSA
mincol=<real>	output only those aligned seqs with >= <real> fraction of column matches
phyla	show the phyla represented in the MSA
pdb	output a cma format file with pdb sequences only
rpdb	remove pdb sequences from the input alignment
rm=<int>	randomly remove all but <int> seqs from an alignment; this can speed up analyses by pruning down a large MSA
U<int>	remove all but one aligned sequence among those sharing ≥ <int> percent identity using an exhaustive algorithm
U	remove identical sequences from a cma file (keeping first occurrences)

MAPGAPS commands. The

Command	code	Description
create	C	Creates mapgaps input files from a CDD hierarchical alignment; also runs S command
Input	I	Creates mapgaps input files from a non-CDD hierarchical aligment
Run	R	Run search & multialign detected sequences
Seed	S	Create, from mapgaps input files, a seed MSA (*.sma) file for a BPPS analysis (see below).

LAPIS. The (preliminary) **LAPIS (Lots of Aligned Proteins Initiated from Scratch)** program takes as input a fasta formatted set of unaligned protein sequences belonging to a specific superfamily and outputs a large MSA. It applies the following steps: (1) Using twkcma, remove redundant sequences from the input set and select a relatively small set of representative sequences. (2) Run GISMO on the representative set to obtain a starting MSA. (3) Create a hiMSA of depth 1 with each sequence serving as a subgroup 'MSA' and a consensus for the MSA as the superfamily template sequence. (4) Run MAPGAPS using the initial hiMSA as the query and the input sequence set as the database to create expanded MSAs for each subgroup. And (5) run MAPGAPS against the full database using the expanded hiMSA as the query to generate the final MSA. Hence, LAPIS provides a way to create very large, relatively accurate MSAs for those superfamilies lacking a manually curated hiMSA. By running the following BPPS program on this MSA, one can obtain a starting, representative hiMSA for manual curation. Note, however, that LAPIS is still under development and is not yet ready for public release.

5. Characterizing protein functional determinants (BPPS)

A large, sufficiently accurate MSA (ideally consisting of at least tens of thousands of sequences) is the first step toward characterizing statistically significant sequence constraints. The next step is to apply **Bayesian Partitioning with Pattern Selection (BPPS)** [19,20] to partition (i.e., subgroup) the MSA into a **hierarchical MSA (hiMSA)(Fig. 6.1A**) with each subgroup defined by 'pattern residues' that best distinguish that subgroup's aligned sequences from sequences in other, closely related subgroups that generally lack that pattern (**Fig. 6.1B**). BPPS defines each subgroup based on statistically significant constrained residues that, presumably, play key roles in subgroup-specific functions. Using MCMC sampling [21] with simulated annealing [22], BPPS [23,24] searches for the mode of the posterior probability distribution over possible hierarchies and, as an aid to biological interpretation, visualizes pattern residues within representative sequences and structures (as illustrated in **Fig. 6.2** below). Among these are residues with well-characterized functions, though our primary focus is on pattern residues of unknown functional relevance—structural interactions among which may suggest hypotheses regarding underlying mechanisms [25-35].

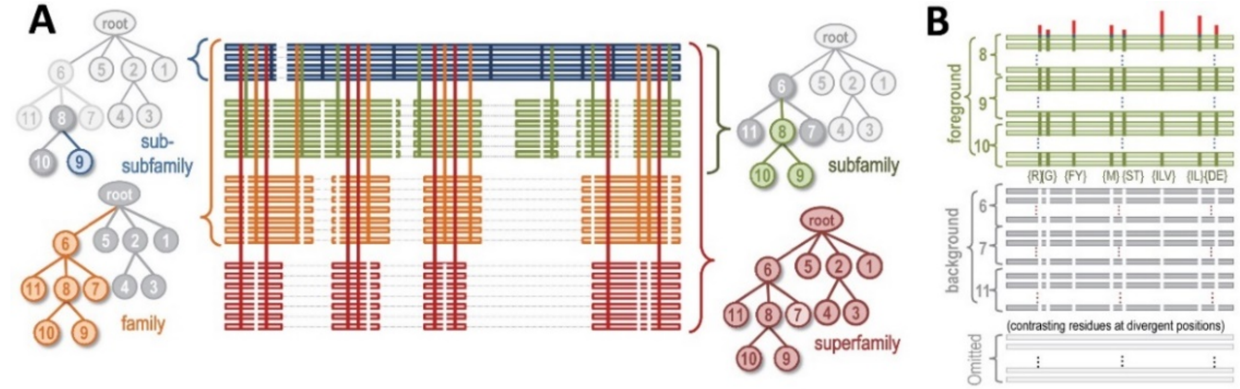
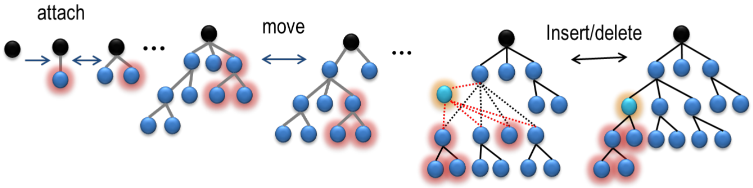


Figure 6.1. Basic features of a protein hiMSA. **A.** Schematic diagram of that part of an hiMSA corresponding to the lineage from the root node to leaf node 9. There is one such diagram for each leaf node in the hierarchy. Horizontal lines represent aligned sequences and are color-coded by level in the hierarchy. Thin light gray horizontal lines represent non-homologous and deleted regions. Vertical lines represent the residue pattern positions upon which the hierarchy is based and are similarly color-coded by levels. On either side are the subtrees for each level of the hierarchy. The colored, gray, and white nodes in each tree correspond, respectively, to their **foreground**, **background**, and non-participating partitions (as explained in panel B). The background for the entire superfamily (not shown) is based on overall amino acid frequencies. **B.** Schematic of a BPPS "**contrast alignment**" corresponding to the (green) subfamily tree rooted at node 8 in panel A. There is one such contrast alignment for each node in a hierarchy. Sequences assigned to node 8's subtree (green nodes of upper right tree in panel A) constitute the 'foreground' partition, those assigned to the rest of node 8's parent subtree (dark gray nodes in panel A) constitute the 'background' partition, and the remaining sequences constitute an omitted, non-participating partition. Horizontal bars represent sequences; these are colored as are the corresponding nodes of the 'subfamily' tree in panel A. Green vertical bars represent the positions of foreground pattern residues (shown below each bar); these diverge from (or contrast with) the background residues at those positions (white vertical bars). Red vertical bars heights quantify the degree of divergence.

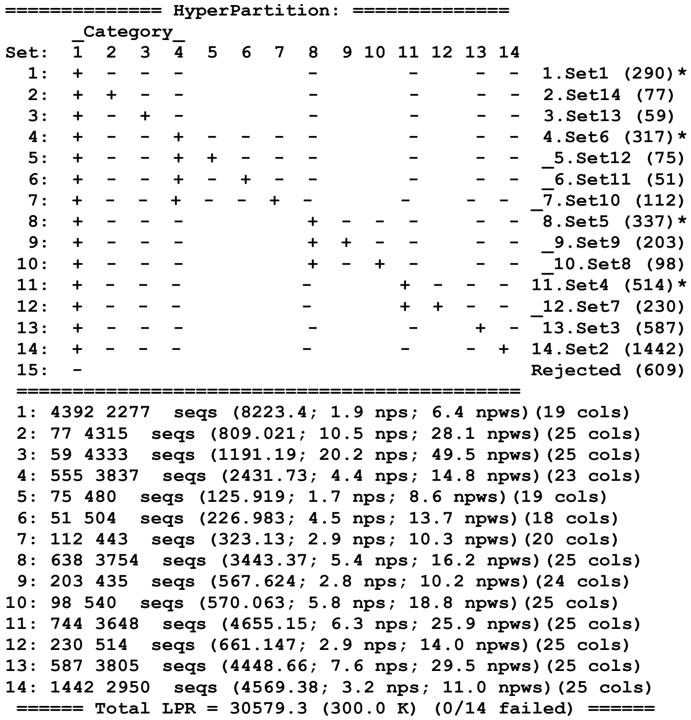
BPPS modes. BPPS runs in various modes as listed in the following table:

Mode	Description
1	Initial hierarchical partitioning of MSA into subgroups
2	Create a hiMSA as in **Fig. 6.1** using a mode 1 checkpoint file; this extends each subgroup alignment
3	Create contrast alignments for a specified node's lineage
A	Run modes 1-3 using default options with optional mapping of pattern residues to structures
Q	Perform a query-centric run in mode 1
E	Evaluate the consistency between two BPPS-generated hiMSAs
H	Run BPPS with a user curated hyperpartition, which need not correspond to a tree, and seed MSAs
M	Show how well various subgroup sequences match each subgroup pattern
V	Create files to identify pattern residues within protein structures for visualization
P	Create PyMOL session files from the files created in mode V

The example shown in **Fig. 6.2** below corresponding to *mode 1*. To obtain the sort of hiMSA shown schematically in **Fig. 6.1**, *modes 2-3* expand each subgroup alignment to include regions corresponding to insertions relative to the common core defined by the root node. BPPS *mode A* runs modes 1-3 in succession. *Mode Q* limits the search to the lineage associated with a user-specified query sequence; this substantially speeds up a search. *Mode E* compares the partitions generated by two distinct runs based on the same input MSA in order to check the consistency between runs. *Mode H* runs BPPS for a fixed set of partitions, termed a hyperpartition (as defined by an *.hpt input file), which are seeded with corresponding seed multiple alignments (as defined by an *.sma input file); this allows construction of an hiMSA that need not correspond to a tree (see below). *Mode M* determines how well the sequences in each partition match each pattern. *Modes V* and *P* are used to generate PyMOL session files showing the locations of pattern residues within available protein structures.



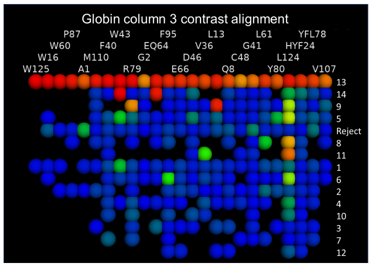
BPPS mode 1 takes as input a multiple sequence alignment (MSA) corresponding to a superfamily of structurally and evolutionarily related protein domains. It explores the 'space' of domain hierarchies by attaching, moving, inserting, and deleting nodes corresponding to functionally divergent subgroups. Each such hierarchy corresponds to a tree, but for reasons explained in the next section, are represented as a **hyperpartition (hpt)**, which (for a tree) consists of an $(n + 1) \times n$ matrix with rows corresponding to nodes in the tree (plus an additional node corresponding to the root node's background set consisting of both shuffled input sequences and 'rejected' sequences lacking superfamily features) and with columns corresponding to contrast alignments (as shown schematically in **Fig. 6.1B**). The figure on the right shows a hpt for globins, where foreground and background partitions are indicated by '+' and '−' symbols, respectively, and omitted partitions are left blank. At the end of each row, the sequence set designation for each node is given with the number of sequences assigned to that set in parentheses. An asterisk at the end of a line indicates an internal (non-leaf) node. The lines below this hpt matrix describe features of each column. Each line lists, for the corresponding contrast alignment, the numbers of foreground and background sequences; the associated **log-probability ratio (LPR)** in nats, in **nats per sequence (nps)** and in **nats per weighted sequence (npws)** (i.e., after down weighting for redundancy); and the number of pattern positions (i.e., of discriminating columns in the contrast alignment). The last line gives the total LPR, the simulated annealing pseudo-temperature, and the fraction of 'failed nodes'—that is, of those nodes lacking statistical significance. The following table lists the most commonly used BPPS mode 1 input options:



```
============== HyperPartition: ==============
     _Category_
Set:  1  2  3  4  5  6  7  8  9 10 11 12 13 14
  1:  +  -  -  -           -        -     -  -  1.Set1 (290)*
  2:  +  +  -  -           -        -     -  -  2.Set14 (77)
  3:  +  -  +  -           -        -     -  -  3.Set13 (59)
  4:  +  -  -  +  -  -  -  -        -     -  -  4.Set6 (317)*
  5:  +  -  -  +  +  -  -  -        -     -  - _5.Set12 (75)
  6:  +  -  -  +  -  +  -  -        -     -  - _6.Set11 (51)
  7:  +  -  -  +  -  -  +  -        -     -  - _7.Set10 (112)
  8:  +  -  -  -           +  -  -  -     -  -  8.Set5 (337)*
  9:  +  -  -  -           +  +  -  -     -  - _9.Set9 (203)
 10:  +  -  -  -           +  -  +  -     -  - _10.Set8 (98)
 11:  +  -  -  -           -        +  -  -  -  11.Set4 (514)*
 12:  +  -  -  -           -        +  +  -  - _12.Set7 (230)
 13:  +  -  -  -           -        -     +  -  13.Set3 (587)
 14:  +  -  -  -           -        -     -  +  14.Set2 (1442)
 15:  -                                        Rejected (609)
 =============================================
 1: 4392 2277  seqs (8223.4; 1.9 nps; 6.4 npws)(19 cols)
 2: 77 4315  seqs (809.021; 10.5 nps; 28.1 npws)(25 cols)
 3: 59 4333  seqs (1191.19; 20.2 nps; 49.5 npws)(25 cols)
 4: 555 3837  seqs (2431.73; 4.4 nps; 14.8 npws)(23 cols)
 5: 75 480  seqs (125.919; 1.7 nps; 8.6 npws)(19 cols)
 6: 51 504  seqs (226.983; 4.5 nps; 13.7 npws)(18 cols)
 7: 112 443  seqs (323.13; 2.9 nps; 10.3 npws)(20 cols)
 8: 638 3754  seqs (3443.37; 5.4 nps; 16.2 npws)(25 cols)
 9: 203 435  seqs (567.624; 2.8 nps; 10.2 npws)(24 cols)
10: 98 540  seqs (570.063; 5.8 nps; 18.8 npws)(25 cols)
11: 744 3648  seqs (4655.15; 6.3 nps; 25.9 npws)(25 cols)
12: 230 514  seqs (661.147; 2.9 nps; 14.0 npws)(25 cols)
13: 587 3805  seqs (4448.66; 7.6 nps; 29.5 npws)(25 cols)
14: 1442 2950  seqs (4569.38; 3.2 nps; 11.0 npws)(25 cols)
 ====== Total LPR = 30579.3 (300.0 K) (0/14 failed) ======
```

Option	**Description**
-heatmap	generate a heatmap for each pattern showing the degree of conservation for each subgroup

-maxcol=<int>	set the maximum number of pattern positions per node to <int> (default: 25)
-minsize=<int>	set the minimum set size to <int> sequences (default: 50)
-minnats=<real>	set the minimum required nats per weighted sequence to <real> (default: 5.0)
-maxdepth=<int>	set the maximum tree depth to <int> (default: 2; though bpps operations may further increase the depth)
-pdb_list=<str>	<str> is a file listing paths to 3D coordinate files; used for creating PyMOL scripts showing constrained residues
-run=P	Create additional output files (requires a checkpoint file created during a previous run)
-seed=<int>	Set the seed for the random number generator to <int>

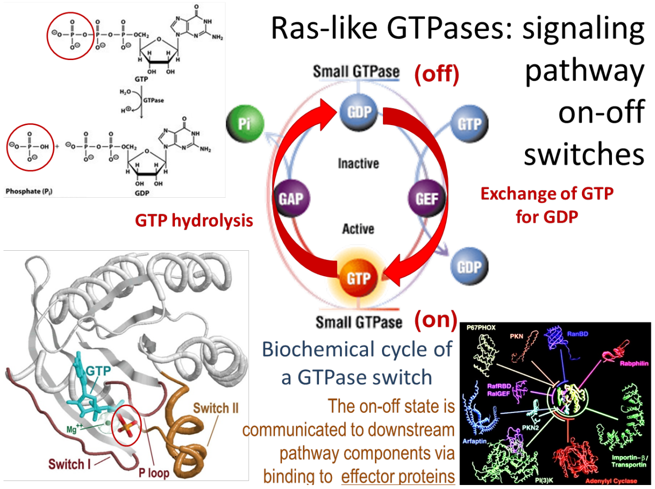


The -heatmap option reveals how strongly pattern residues are conserved in the foreground and in other partitions for each contrast alignment (see image on the right). This option, which requires that PyMOL be installed and on your path, reveals pattern residues that are 'cross conserved'; that is, are also conserved in some of the background or omitted subgroups. If so, then performing a second analysis without those subgroups harboring cross-conserved residues may more clearly define the functionally important residues for the foreground subgroup. Alternatively, a user may define a new hyperpartition that does not correspond to a tree and therefore better models the observed patterns of conservation; this is discussed and illustrated below using BPPS in the H mode.

Considerations when performing BPPS. It is important to understand the various factors influencing BPPS. For example, note that the contrasting residues for each node are determined by both the foreground and background aligned sequence sets. Factors having a minor influence on results include the degree of sequence redundancy that remains after down weighting, the number of aligned columns in the MSA, and the choice of various sampling parameter settings (e.g., allowed pattern residues, priors, convergence criteria, etc.). Also, very small subgroups may not have enough members to reach statistical significance and, consequently, the sampler may include that subgroup in a related subgroup from which it has diverged—thereby obscuring the latter subgroup's distinctive features. Another source of subgroup assignment ambiguity is the inclusion of sequences that correspond to pseudogenes or sequencing errors or that were translated from DNA for organisms whose genetic code was incorrectly assigned. During BPPS sampling over a given hierarchy (done using the -run=o option) or over alternative hierarchies such sequences might continuously be reassigned to alternative subgroups due to inherent ambiguities. Running **BPPS in the 'E' mode** can help distinguish between sequences that are consistently sampled to a specific subgroup and those that are sampled across various subgroups; the latter may then be eliminated from the main alignment. (BPPS-E will create an MSA of consistently sampled sequences as input to a subsequent BPPS analysis to thereby eliminate ambiguous sequences.) Typically insert regions are not included in a BPPS mode '1' analysis, though running BPPS in modes '2' and '3' aims to add these regions after the fact. Moreover, for BPPS mode '1' the relationships among sequence sets are assumed to be hierarchical, which may not accommodate their actual relationships. For these reasons, it is often helpful to exercise more control over an analysis by running **BPPS in the 'H' mode**—as described and illustrated below. The interpretation of results is facilitated by the availability of 3D structures, the inclusion of other, interacting subunits, and published information).

- Predictive probabilities and sequencing errors? Remove ambiguous sequences for clarity?
    - ▪ Compute intersection of sequence sets? (Same input set; use set_typ).
    - o Optimize over sequences only (new -run=S or O option?); then run mode 'E'.
- Repeats and isolated conserved motifs within sequences and their constraints; new BG set. -mtf=option

BPPS analysis of Ras-like GTPases. To illustrate BPPS we apply it to phosphate-binding loop (P-loop) GTPases [36], which bind to GDP or GTP via Walker A (G-K-[ST]) residues, which corresponds to the P-loop, and Walker B (D-x-x-G) residues, the conserved aspartate (D) of which interacts (indirectly through a water molecule) with the Mg^{++} ion that coordinates with nucleotide phosphate groups. An important subgroup of P-loop GTPases are Ras-like GTPases, which includes Ras, Rab, Rho/Rac, Ran, Arf, and Arf-like (Arl) GTPases and α subunits of heterotrimeric G proteins. Ras-like GTPases [37] function as on-off switches within eukaryotic signaling pathways regulating diverse cellular processes, including vesicle transport, embryonic development, the sensation of vision, odor, taste and pain, microtubule assembly and cell division. These GTPases are associated both with guanine nucleotide exchange factors (GEFs), which turn them on by mediating the exchange of GTP for GDP, and with GTPase activating proteins (GAPs), which turn them off by stimulating hydrolysis of GTP to GDP. The on- or off-state of Ras-like GTPases is communicated to effector proteins through conformational changes within their switch I and II regions, which detect the presence or absence of the γ phosphate of bound guanine nucleotide.



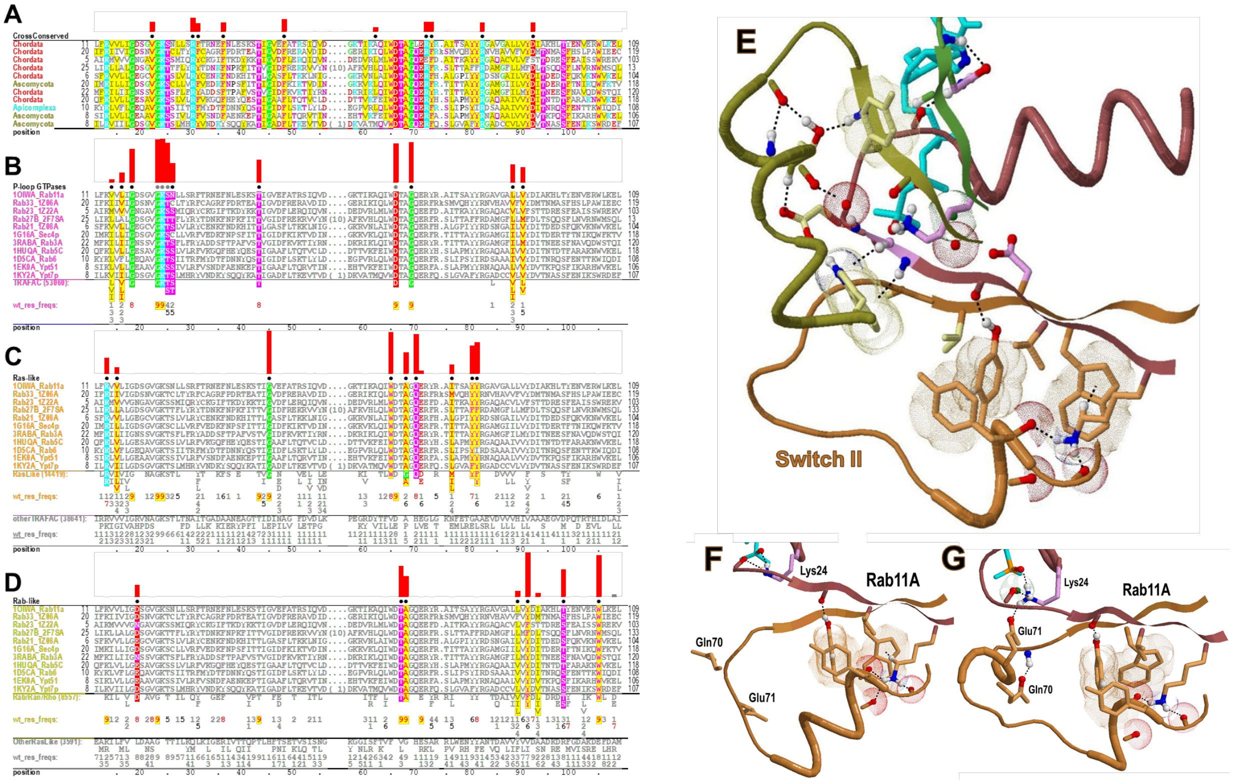
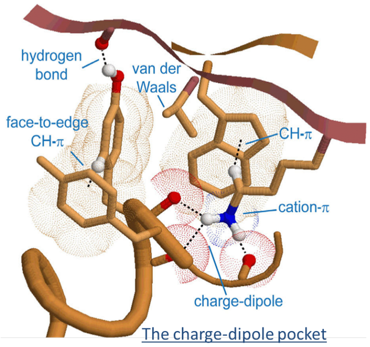


Figure 6.2. BPPS-1 analysis of Ras-like GTPases. **A-D**. Four contrast alignments highlighting pattern residues that most distinguish the subgroups to which Rab GTPases belong [34,35]. **A.** Alignment highlighting all residues that are conserved among the 11 aligned sequences representative of Rab-like GTPases. The bullets above specific columns in the alignment indicate residue positions that are well conserved within the displayed sequences but that are not distinctive of the subgroups corresponding to the B, C and D contrast alignments; the heights of the red bars above these positions indicate (on a semi-logarithmic scale) the degree of conservation relative to standard amino acid background frequencies. The leftmost column gives the phylum to which each

sequence belongs and is color coded by kingdom (metazoans, red; protozoans, cyan; fungi, dark yellow; plants, green; bacteria, purple; archaea, blue). **B.** Contrast alignment highlighting pattern positions distinctive of P-loop GTPases, with corresponding pattern residues indicated below the alignment; directly below these, corresponding residue frequencies are given in integer tenths. A '7', for example, indicates that the corresponding residue occurs in 70-80% of the sequences. Above highlighted columns are red bars quantifying the selective constraints imposed on pattern positions. The leftmost column gives the protein identifiers for the 11 representative sequences shown and below this column is given the name of the foreground set and (in parentheses) the total number of sequences assigned to the foreground. **C**. Highlighted residues distinguishing Ras-like GTPases from other P-loop GTPases [36]. The format used is as described for alignment B, except that corresponding information regarding the background sequence set is also provided. **D**. Residues distinguishing the **Rab-like subfamily** from other Ras-like GTPases. **E**. Structural locations of subgroup-specific residues within Rab11A bound to a GTP analog [38]. Residues generally conserved in all P-loop GTPases (magenta sidechains) bind to GTP or to a GTP-bound Mg^{++} ion. Residues with orange and yellow sidechains correspond to Ras-like family and Rab-like subfamily GTPases—that is, to pattern residues highlighted in C and D, respectively. Five of the Ras-like residues occur in the Switch II region that undergoes conformational changes associated with signal transduction. These five Ras-like residues [35] mutually-interact near the C-terminal end of the switch II region. As described below, the Rab-like subfamily residues (yellow side-chains) form aromatic CH-π interactions hypothesized to stabilize glycine 'flexible hinges' within nucleotide binding loops, thereby facilitating nucleotide binding and release [34]. In crystal structures these form distinct Rab11A conformations (**F**, **G**) [39] that have been proposed to play a role in the switching mechanism [35] via repositioning of two other Ras-like residues (Gln70 and Glu71 of Rab11A) involved in GTP hydrolysis [40] and in nucleotide exchange [41,42], respectively.

Shown in **Fig. 6.2** are four BPPS-generated contrast alignments corresponding to the lineage for Rab-like subfamily, which consists of Rab, Rho and Ran GTPases. (Note that, for clarity, subgroups containing cross-conserved pattern residues were omitted from this analysis.) The highlighted positions within aligned Rab-like sequences reveal discriminating features distinctive of P-loop GTPases [35] (**Fig. 6.2B**), of the Ras-like family [34] (**Fig. 6.2C**), and of the Rab-like subfamily (**Fig. 6.2D**)—as well as other conserved residues that are not distinctive of any of these categories (**Fig. 6.2A**). Roles for several of the Ras-like GTPase pattern residues (**Fig. 6.2C**) in GTP hydrolysis and nucleotide exchange were previously proposed. However, BPPS identifies five additional Ras-like residues: (i) the pattern [RK]-x-[ILV] preceding the P-loop, (ii) the pattern [WF] directly preceding the Walker B aspartate, and (iii) the pattern [YF]-[YF] at the C-terminal end of the switch II region. These residues form an aromatic pocket around the negative-dipole moment at the end of the switch II helix with the positively charged pattern residue inserted into the pocket. This helix is oriented in a specific direction away from the GTPase core but is reoriented upon rearrangement of these residues, which were collectively termed the **charge-dipole pocket** (**Figs 6.2F,G**). The charge-dipole pocket occurs in both the on and off states and both this configuration and an alternative configuration occur within the unit cell of a single crystal structure of Rab5a GTPase in the off state. Thus, the charge-dipole pocket configuration is closely associated, not with the on or off state, but rather with formation of an outward-oriented helix and, as a result, with restructuring of the switch II N-terminal region, which plays a critical role both in sensing the on/off state and in mediating GTP hydrolysis and nucleotide exchange via two other Ras-like residues {ref}, which correspond to Gln70 and Glu71 in Rab11A (**Figs 6.2F,G**).



The charge-dipole pocket

An unusual homodimeric Rab27 configuration (**Fig. 6.3A**), in which the switch I, switch II and inter-switch regions of each subunit are exchanged and protrude out like an antenna, suggests a role for the swII-CT residues: In this configuration both subunits form charge-dipole pockets (**Fig. 6.3B**), which thus are structurally compatible with the outward-directed switch II helices. An association between the charge-dipole pocket and such switch II restructuring is suggested by comparisons between this unusual, homodimeric form of GDP-bound Rab27 and monomeric forms of Rab family GTPases. In the homodimeric form, the switch II region forms a long α helix that is directed away from the structural core of one subunit and toward the structural core of the other subunit. Presumably this helix lacks the conformational strain typically imposed on monomeric GTPases—the switch II regions of which need to bend around and reconnect to the structural core. Hence the charge-dipole pocket is structurally compatible with formation of this unusual, outward-directed switch II helix. Likewise, within monomeric Ras-like GTPases formation of a charge-dipole pocket is associated with an outward-directed switch II helix (**Fig.

6.3C), whereas its disruption is associated with disruption of this outward-directed helix (**Fig. 6.3D**)—a theme that is repeated in other Ras-like GTPases {ref}.

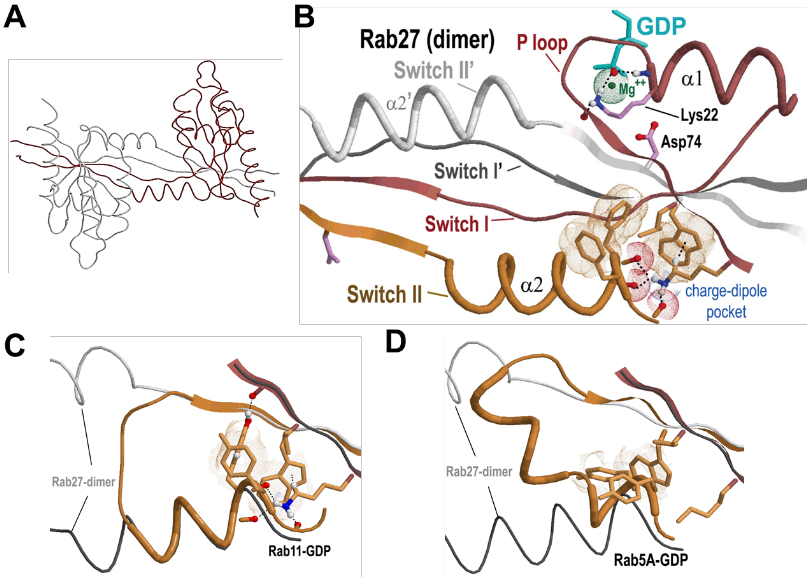
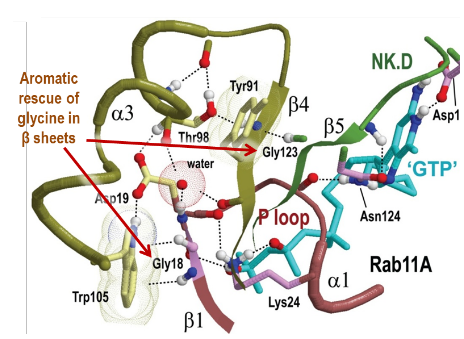


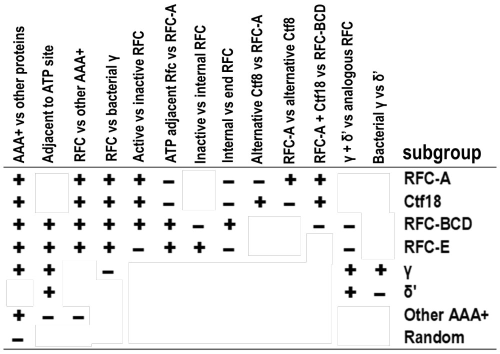
Figure 6.3. Insight from an unusual homodimeric configuration of Rab27. **A.** In the Rab27 homodimer [43], the switch I, switch II and inter-switch regions of each subunit are exchanged and protrude out like an antenna. **B.** In this configuration both subunits form a charge-dipole pocket. **C.** In the monomeric Rab11 GTPase formation of the charge-dipole pocket is also associated with an outward-directed switch II helix that traces out the same path as the homodimeric helix. **D.** In the monomeric Rab5 GTPase disruption of the charge-dipole pocket is associated with disruption of the outward-directed switch II helix. This correlation is seen for various Ras-like GTPases, including Arf, Arl, Sar, and Gα GTPases [35].



BPPS also identifies (**Fig. 6.2D**) both two Rab-like pattern residues (forming a 'T-A' motif) previously proposed to perform a role in nucleotide exchange [40] and four other previously unidentified Rab-like residues forming a structural component, termed the glycine brace [34]. These include an aromatic residue that forms a CH-π interaction with a Ras-like conserved glycine at the start of the guanine-binding loop (Gly123; see figure on the right) and a second aromatic residue (nearly always a tryptophan) that forms CH-π and NH-π interactions with a conserved glycine at the start of the phosphate-binding P-loop (Gly18 in the figure). Such aromatic interactions are believed to stabilize the β-strand conformation of glycine [44]. The two other Rab-like residues (typically an aspartate and a serine or threonine), together with a conserved buried water molecule, form a network of interactions connecting the two aromatic residues. These observations suggest that the two glycine residues serve as hinges for the P-loop and for the guanine-binding loop and that the glycine brace facilitates guanine nucleotide binding and release by either interacting with or dissociating from these glycine hinges. Consistent with this notion, these aromatic-glycine interactions are disrupted in the structure of Ran GTPase bound to its exchange factor RCC1 (pdb_id: 1i2m) [45] (not shown here).

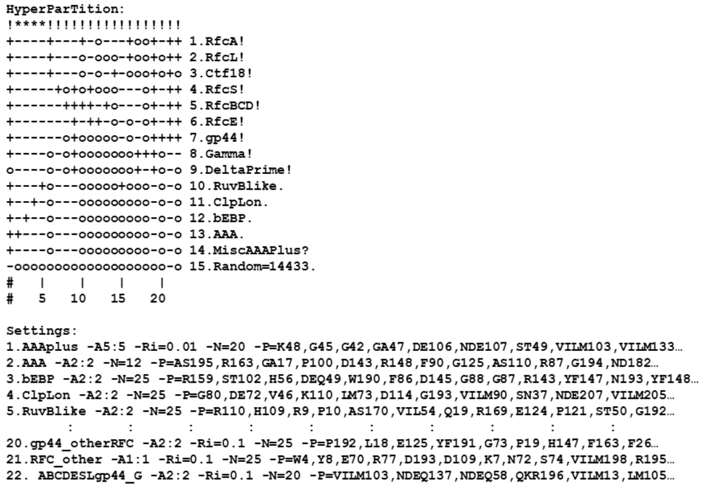
cccd

BPPS H-mode. Protein superfamilies can exhibit complex patterns of residue constraints that cannot be modeled hierarchically (i.e., as a tree). This can occur when one or more protein subgroups inherit certain biochemical properties from a common ancestor that, presumably due to a relaxation of selective constraints, are lost in other subgroups that have descended from the same ancestor. This is seen, for example, for certain families of AAA+ ATPases, which mediate diverse cellular activities, including membrane fusion, DNA clamp loading and replication, microtubule dynamics, intracellular transport, transcriptional activation, protein refolding or degradation, and the assembly and disassembly of protein complexes. For **eukaryotic replication factor C (RFC)** DNA clamp loader AAA+ domains [19], subunits B, C and D share constraints both with RFC-A (because these all hydrolyze ATP) but not with inactive RFC-E, and with RFC-E (because these all trans-activate ATP hydrolysis) but not with RFC-A (which does not). Moreover, all RFC subunits share constraints both in common with and distinct from bacterial clamp loader γ subunits. To model such complex relationships, the BPPS H-mode may be used, the output of which is illustrated for RFC subunits in **Fig. 6.4**. This requires as input both a user-defined hyperpartition, which may be edited using the **edit_hpt** program, and a representative 'seed' alignment for each subgroup. The user-defined hpt file can include specific paramaters for each contrast alignment. The H-mode can also rerun an analysis using a larger (updated) MSA of the same superfamily and of the same length as a previous analysis. The format required for an input hyperpartition (hpt) is illustrated and described in **Fig. 6.5**; a corresponding output hpt is shown in **Fig. 6.6**. Running BPPS in H-mode creates both output contrast MSAs showing the locations of constrained residues and, if structural coordinates are available, PyMOL session files showing the structural locations of subgroup-specific residues. These output files are illustrated in **Figs 6.7** and **6.8**.



AAA+ vs other proteins	Adjacent to ATP site	RFC vs other AAA+	RFC vs bacterial γ	Active vs inactive RFC	ATP adjacent Rfc vs RFC-A	Inactive vs internal RFC	Internal vs end RFC	Alternative Ctf8 vs RFC-A	RFC-A vs alternative Ctf8	RFC-A + Ctf18 vs RFC-BCD	γ + δ' vs analogous RFC	Bacterial γ vs δ'	subgroup
+		+	+	+	−		−	−	+	+			RFC-A
+		+	+	+	−		−	+	−	+			Ctf18
+	+	+	+	+	+	−	+			−	−		RFC-BCD
+	+	+	+	−	+	+	−				−		RFC-E
+	+		−								+	+	γ
	+										+	−	δ'
+	−	−											Other AAA+
−													Random

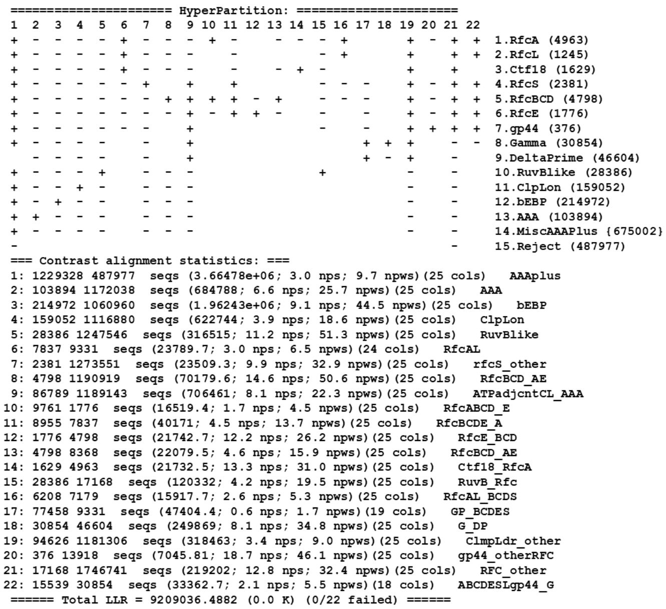
Fig. 6.4. Hyperpartition for characterizing non-hierarchical BPPS constraints imposed on DNA clamp loader AAA+ domains.



```
HyperParTition:
!****!!!!!!!!!!!!!!!!
+----+---+-o---+oo+-++ 1.RfcA!
+----+---o-ooo-+oo+o++ 2.RfcL!
+----+---o-o-+-ooo+o+o 3.Ctf18!
+-----+o+o+ooo---o+-++ 4.RfcS!
+------++++-+o---o+-++ 5.RfcBCD!
+-------+-++-o-o-o+-++ 6.RfcE!
+------o+ooooo-o-o++++ 7.gp44!
+----o-o+ooooooo+++o-- 8.Gamma!
o----o-o+ooooooo+-+o-o 9.DeltaPrime!
+---+o---ooooo+ooo-o-o 10.RuvBlike.
+--+-o---ooooooooo-o-o 11.ClpLon.
+-+--o---ooooooooo-o-o 12.bEBP.
++---o---ooooooooo-o-o 13.AAA.
+----o---ooooooooo-o-o 14.MiscAAAPlus?
-ooooooooooooooooooo-o 15.Random=14433.
#   |    |    |    |
#   5   10   15   20

Settings:
1.AAAplus -A5:5 -Ri=0.01 -N=20 -P=K48,G45,G42,GA47,DE106,NDE107,ST49,VILM103,VILM133…
2.AAA -A2:2 -N=12 -P=AS195,R163,GA17,P100,D143,R148,F90,G125,AS110,R87,G194,ND182…
3.bEBP -A2:2 -N=25 -P=R159,ST102,H56,DEQ49,W190,F86,D145,G88,G87,R143,YF147,N193,YF148…
4.ClpLon -A2:2 -N=25 -P=G80,DE72,V46,K110,LM73,D114,G193,VILM90,SN37,NDE207,VILM205…
5.RuvBlike -A2:2 -N=25 -P=R110,H109,R9,P10,AS170,VIL54,Q19,R169,E124,P121,ST50,G192…
       :       :       :       :       :       :       :       :       :       :
20.gp44_otherRFC -A2:2 -Ri=0.1 -N=25 -P=P192,L18,E125,YF191,G73,P19,H147,F163,F26…
21.RFC_other -A1:1 -Ri=0.1 -N=25 -P=W4,Y8,E70,R77,D193,D109,K7,N72,S74,VILM198,R195…
22. ABCDESLgp44_G -A2:2 -Ri=0.1 -N=20 -P=VILM103,NDEQ137,NDEQ58,QKR196,VILM13,LM105…
```

Figure 6.5. Input hyperpartition (hpt) illustrated for AAA+ RFC proteins. There are two sections: the first section (labeled 'HyperParTition") shows the columns and rows of the hpt. The first row consists of two types of characters ('!', and '*') corresponding to the columns of the hpt, where '!' and '*' indicates that a contrast alignment for that column should or should not be printed out. The rows of the hpt correspond to the subgroups being modeled. Each row consists of a series of '+', '-', and 'o' characters, which indicate that the subgroup in that row is assigned to the foreground, background, or omitted partition for that column's contrast alignment. This string is followed by the row number, the name of the subgroup, and a '!', '.', or '?' character, where a '!' and '.' indicates that the contrast alignments for which that subgroup is in the foreground should or should not be printed, respectively; and where a '?' indicates that the row does not correspond to a specific subgroup (and thus that the contrast alignment should not be printed). The second section (labeled as 'Settings') lists the names of the contrast alignment for each column along with (optionally) parameter settings specific to each column; optionally, this includes a seed pattern.



```
===================== HyperPartition: =====================
1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
+  -  -  -  -  +  -  -  -  +  -     -  -  -  +        +  -  +  +  1.RfcA (4963)
+  -  -  -  -  +  -  -  -     -           -  +        +     +  +  2.RfcL (1245)
+  -  -  -  -  +  -  -  -     -     -  +  -           +     +     3.Ctf18 (1629)
+  -  -  -  -  -  +     +     +           -  -  -     +  -  +  +  4.RfcS (2381)
+  -  -  -  -  -  -  +  +  +  +  -  +     -  -  -     +  -  +  +  5.RfcBCD (4798)
+  -  -  -  -  -  -  -  +  -  +  +  -     -     -     +  -  +  +  6.RfcE (1776)
+  -  -  -  -  -  -     +                 -     -     +  +  +  +  7.gp44 (376)
+  -  -  -  -     -     +                    +  +  +     -  -  8.Gamma (30854)
   -  -  -  -     -     +                    +  -  +     -     9.DeltaPrime (46604)
+  -  -  -  +     -  -  -                 +           -     -     10.RuvBlike (28386)
+  -  -  +  -     -  -  -                             -     -     11.ClpLon (159052)
+  -  +  -  -     -  -  -                             -     -     12.bEBP (214972)
+  +  -  -  -     -  -  -                             -     -     13.AAA (103894)
+  -  -  -  -     -  -  -                             -     -     14.MiscAAAPlus {675002}
-                                                           -     15.Reject (487977)
=== Contrast alignment statistics: ===
 1: 1229328 487977  seqs (3.66478e+06; 3.0 nps; 9.7 npws)(25 cols)   AAAplus
 2: 103894 1172038  seqs (684788; 6.6 nps; 25.7 npws)(25 cols)   AAA
 3: 214972 1060960  seqs (1.96243e+06; 9.1 nps; 44.5 npws)(25 cols)   bEBP
 4: 159052 1116880  seqs (622744; 3.9 nps; 18.6 npws)(25 cols)   ClpLon
 5: 28386 1247546  seqs (316515; 11.2 nps; 51.3 npws)(25 cols)   RuvBlike
 6: 7837 9331  seqs (23789.7; 3.0 nps; 6.5 npws)(24 cols)   RfcAL
 7: 2381 1273551  seqs (23509.3; 9.9 nps; 32.9 npws)(25 cols)   rfcS_other
 8: 4798 1190919  seqs (70179.6; 14.6 nps; 50.6 npws)(25 cols)   RfcBCD_AE
 9: 86789 1189143  seqs (706461; 8.1 nps; 22.3 npws)(25 cols)   ATPadjcntCL_AAA
10: 9761 1776  seqs (16519.4; 1.7 nps; 4.5 npws)(25 cols)   RfcABCD_E
11: 8955 7837  seqs (40171; 4.5 nps; 13.7 npws)(25 cols)   RfcBCDE_A
12: 1776 4798  seqs (21742.7; 12.2 nps; 26.2 npws)(25 cols)   RfcE_BCD
13: 4798 8368  seqs (22079.5; 4.6 nps; 15.9 npws)(25 cols)   RfcBCD_AE
14: 1629 4963  seqs (21732.5; 13.3 nps; 31.0 npws)(25 cols)   Ctf18_RfcA
15: 28386 17168  seqs (120332; 4.2 nps; 19.5 npws)(25 cols)   RuvB_Rfc
16: 6208 7179  seqs (15917.7; 2.6 nps; 5.3 npws)(25 cols)   RfcAL_BCDS
17: 77458 9331  seqs (47404.4; 0.6 nps; 1.7 npws)(19 cols)   GP_BCDES
18: 30854 46604  seqs (249869; 8.1 nps; 34.8 npws)(25 cols)   G_DP
19: 94626 1181306  seqs (318463; 3.4 nps; 9.0 npws)(25 cols)   ClmpLdr_other
20: 376 13918  seqs (7045.81; 18.7 nps; 46.1 npws)(25 cols)   gp44_otherRFC
21: 17168 1746741  seqs (219202; 12.8 nps; 32.4 npws)(25 cols)   RFC_other
22: 15539 30854  seqs (33362.7; 2.1 nps; 5.5 npws)(18 cols)   ABCDESLgp44_G
 ====== Total LLR = 9209036.4882 (0.0 K) (0/22 failed) ======
```

Figure 6.6. Output hyperpartition (hpt) for AAA+ RFC proteins. Based on the specifications provided by the input hpt and an input MSA, BPPS returns an optimized partitioning based on underlying constraints. The output again has two sections: The first section (labeled 'HyperPartition") shows the columns and rows of the hpt. The rows again correspond to the subgroups, where each row consists of a series of '+', '-', and ' ' characters, corresponding to the foreground, background, or omitted partition for that column's contrast alignment, followed by the row number, the subgroup name and (parenthetically) the number of sequences assigned to that subgroup. The second section lists statistics for the contrast alignments corresponding to the columns of the hpt. The information in each of these rows consists of: the column number; the number of foreground and background sequences; the log probability ratio (LPR) in nats; the nats per sequence (nps); the nats per weighted sequence (npws); the number of pattern positions (i.e., discriminating columns in the MSA); and the name of the contrast alignment. The last line gives the total LPR (in nats); the final sampling 'temperature'; and the number of contrast alignments that failed (i.e., that lack statistical significance).

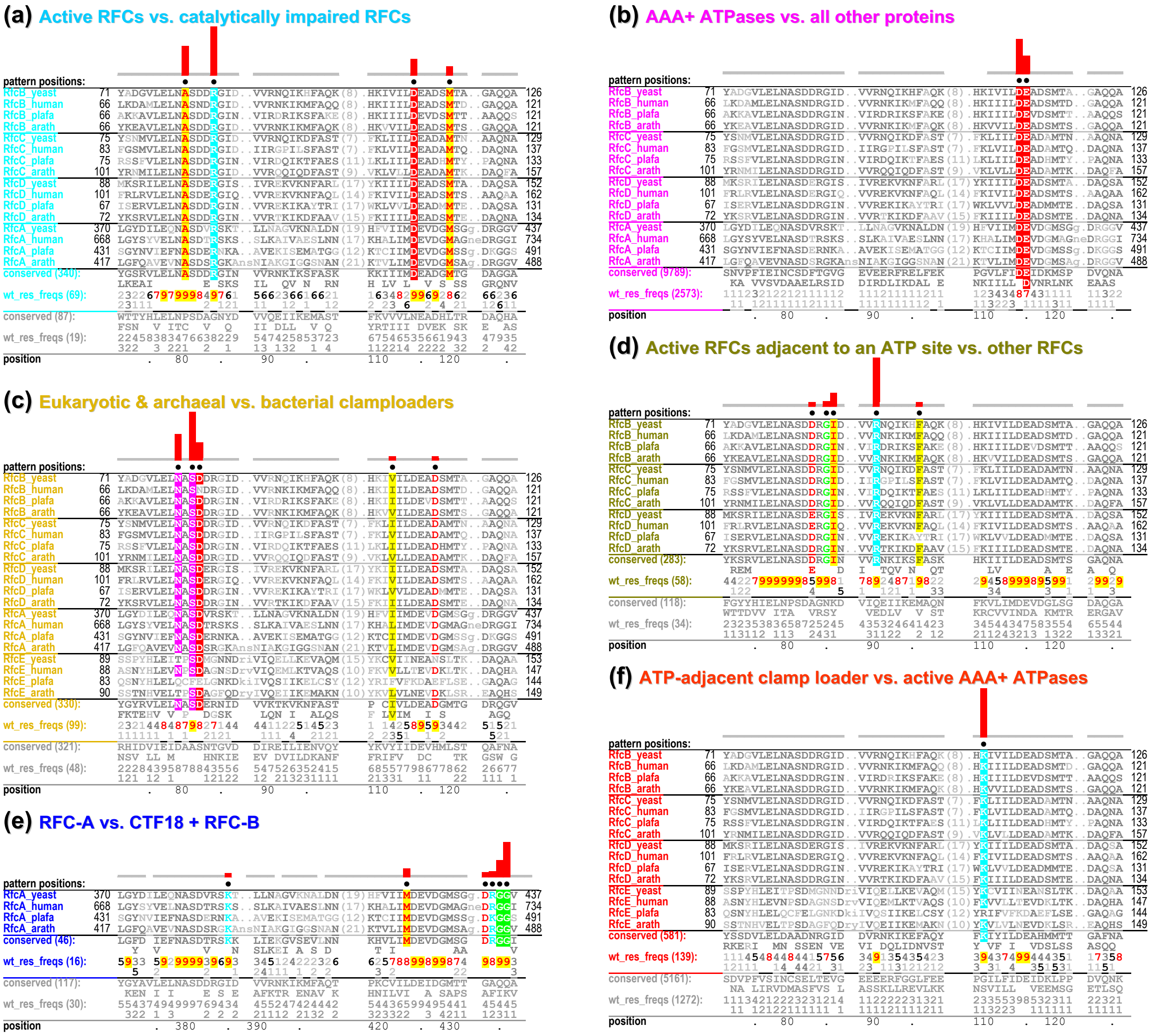


Fig. 6.7. BPPS H-mode generated contrast alignments showing sequence constraints most distinctive of various types of replication factor C (RFC) clamp-loader subunits [46]. The sequences shown span a functionally crucial region of RFC subunits, within which most of the strongest constraints occur. Each alignment is highlighted to reveal constraints distinguishing a specific subgroup of RFC subunits from closely related subgroups. The alignments include representative RFC subunits from fungi (budding yeast), animals (human), protozoans (malaria parasite) and plants (mouse-ear cress). As for **Fig. 6.2**, the conserved patterns and corresponding frequencies that are distinctive of the foreground are shown directly below each alignment and, below this, the conserved patterns distinctive of the background are shown in gray. **(a)** Constraints most distinguishing active RFC ATPases from catalytically inactive RFC-E subunits. **(b)** Constraints most distinguishing active AAA+ ATPases from all other proteins. (Background pattern residues and frequencies are not shown.) **(c)** Constraints most distinguishing eukaryotic and archaeal clamp-loader RFC subunits from bacterial clamp-loader γ subunits. **(d)** Constraints most distinguishing active RFC ATPases that interact with the ATP site of an adjacent RFC subunit from other RFC subunits. **(e)** Constraints most distinguishing the large RFC subunit (RFC-A) from both an alternative large RFC subunit (CTF18) [47] and the structurally adjacent small RFC subunit (RFC-B). **(f)** Constraint most distinguishing all (eukaryotic, archaeal, bacteriophage and bacterial) clamp-loader ATPases interacting with an adjacent ATP site from other AAA+ ATPases.

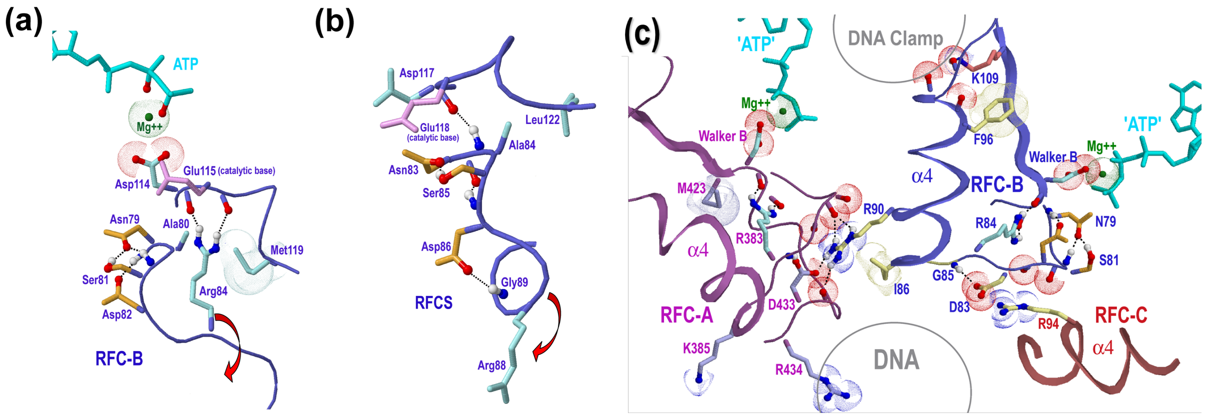


Figure 6.8. Structural features associated with the six categories of RFC constraints shown in **Fig. 6.7**. The central hole of the RFC complex, through which DNA presumably is thread, is located at the bottom of each panel. Oxygen, nitrogen and (predicted) hydrogen atoms establishing hydrogen bonds (broken lines) or involved in ionic interactions are shown in red, blue and white, respectively. Ionic and van der Waals interactions are shown as dot clouds. Residues with cyan side chains distinguish active RFC ATPases from catalytically impaired RFC-E subunits; the magenta residue corresponds to the putative catalytic base shared by all active AAA+ ATPases; orange residues distinguish all RFC subunits from bacterial clamp-loader ATPases; and yellow residues distinguish active RFC ATPases that interact with an adjacent ATP site from other RFC subunits. **(a)** Conformation of RFC-B when bound to ATP and the clamp [5]. In all four active RFC ATPases, the arginine corresponding to Arg84 in RFC-B forms hydrogen bonds with main-chain oxygen atoms on either side of the putative catalytic base, as shown. **(b)** A conformation of ADP-bound RFCS, the archaeal small subunit corresponding to RFC-B [31]. The Walker B-interacting arginine (Arg84 in RFC-B or Arg88 in RFCS) is repositioned and, based on a model of the RFC–DNA–clamp complex, could interact with DNA thread through the clamp. **(c)** Regions of interaction between RFC-A, RFC-B and RFC-C subunits in the crystal structure of the RFC–ATP– clamp complex. The residue side chains corresponding to the NxSD motif in RFC-A have been omitted for clarity. The phenylalanine in RFC-B (Phe96) seems to form a hydrophobic pocket for Lys109. Blue residues distinguish RFC-A from other RFC subunits; the red residue (K109) most clearly distinguishes all clamp-loader subunits interacting with an adjacent ATP site from other AAA+ ATPases.

The get_pdb, twkpdb, and **vsi2pml (after bpps V) programs**. For BPPS and several other programs to map the structural locations of pattern residues (as in **Fig. 6.8**) requires as input structural coordinate files, which may be retrieved using the **get_pdb** program; **get_pdb** also calls the reduce program {ref} to add modeled hydrogen atoms to identify hydrogen bonds automatically based on geometric criteria. **twkpdb** can be used to analyze or modify (i.e., ‘tweak’) a pdb-formated coordinate file. To conserved disc space, BPPS may store pattern residue locations for various proteins of known structure within a *.vsi (visualize structural interactions) file, from which the **vsi2pml** program generates corresponding PyMOL session (*.pse) files, automatically. See the Auxilliary Programs section below for detailed descriptions.

6. Correspondence between sequence and structural constraints (SIPRIS)

SIPRIS (Structurally Interacting Pattern Residues' Inferred Significance) identifies *correlations between BPPS- and structurally-defined residue sets*. SIPRIS takes as input a set of BPPS-defined pattern residues and corresponding structural coordinates for proteins of interest and, for each structure, identifies statistically significant clusters or hydrogen bond networks of pattern-residues. The simplest cluster is predefined, for instance, by residues at the interface between subunits or in contact with substrate, as illustrated in **Fig. 7.1** for *N*-acetyltransferases, which transfer an acetyl group from CoA to a target substrate.

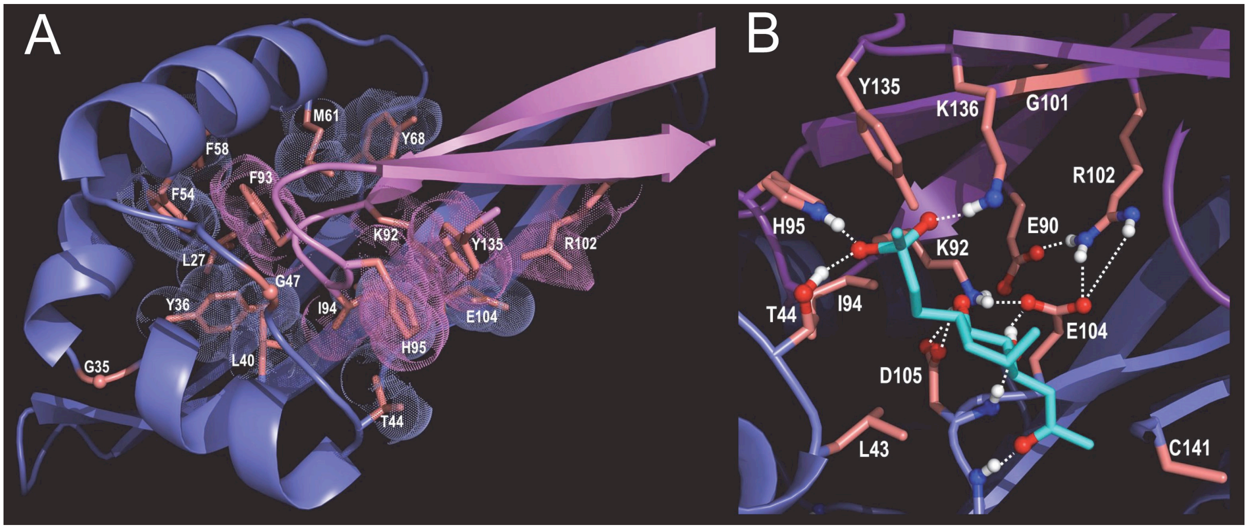


Figure 7.1. SIPRIS [48,49] estimates the statistical significance of all but two of the 25 residues most distinctive of the glucosamine-6-phosphate *N*-acetyltransferase (Gna1) family; these either (**A**) occur at the interface between homodimeric subunits (blue & pink cartoon traces) (***p* = 8.5×10⁻⁷**) or (**B**) interact with substrate (shown in cyan) (***p* = 6.8×10⁻⁵**) (pdb_id: 4ag9 [50]). The remaining two residues (not shown) are: Lys116, which may position CoA for catalysis by interacting with a CoA phosphate group, and Cys141, which covalently links to the sulfur atom of CoA within Gna1 [50].

SIPRIS can also select the optimal, structurally-defined cluster among a nested set of clusters: Starting from the highest scoring BPPS pattern residue or from a user-defined residue, molecule or atom, SIPRIS first creates an initial cluster by sequentially adding "structurally-adjacent" residues until reaching a pattern residue. (Structural adjacency can be defined based either on the closest hydrogen bond to a current member of the cluster or on the distance either from the starting residue or from another current member of the cluster.) Next, it adds more residues in this way until reaching the next pattern residue; this is the second cluster. Finally, it repeats this process until generating a cluster containing all BPPS residues. From this nested set, SIPRIS selects the cluster that most significantly overlaps with the BPPS-defined residue set (after adjusting for the number of hypotheses considered); this is illustrated for Rab, Rho and Ran (R^3) GTPases in **Fig. 7.2**.

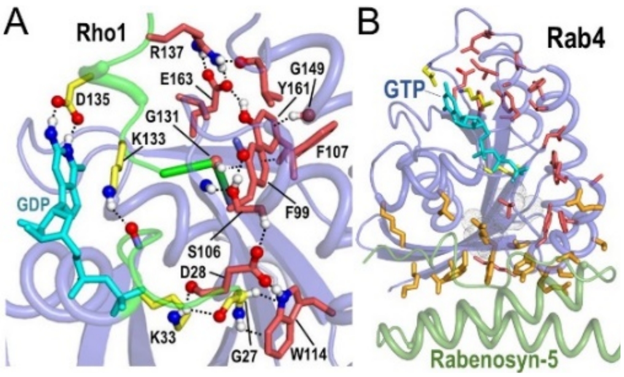


Fig. 7.2. SIPRIS defined GTPase clusters. **A**. Rab/Rho/Ran (R^3) hydrogen-bond network (red sidechains) in Rho1 (pdb_id: 3refB [51]; ***p* = 6.2×10⁻⁵**). This includes a salt bridge (R137-E163) and CH-π interactions (G27-W114; G131-F99) hypothesized to modulate nucleotide exchange by stabilizing the P-loop and guanine binding loop [20,34,44]. These loops (bright green backbone traces) harbor residues (yellow sidechains) that bind to guanine-nucleotide and that are distinctive of all GTPases. **B**. A Rab4 R^3 hydrogen bond network (red sidechains) (***p* = 2.6×10⁻⁶**) and a Rab-specific cluster (orange sidechains) (***p* = 2.9×10⁻⁸**) contacting the Rabenosyn-5 effector (pdb_id: 1z0kA [52]).

A BPPS-SIPRIS comparative analysis of exonuclease-endonuclease phosphatases (EEPs) [53] identifies residues presumably responsible for the functional specificity of APE1 endonucleases (**Fig. 7.3**) and of inositol polyphosphate 5-phosphatases (INPP5) (**Fig. 7.4**), which cleave phosphodiester bonds in nucleic acids and phospholipids, respectively. These analyses suggest that the same core structure and catalytic residues can mediate very different reactions by interacting with a network of family-specific residues, thereby presumably forming a substrate-specific 'reaction chamber'.

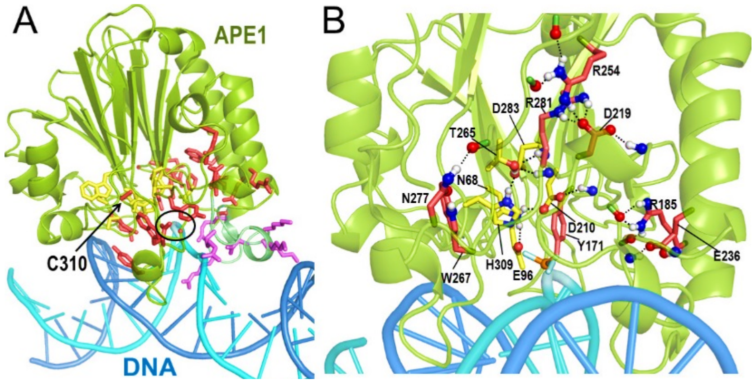


Fig. 7.3. BPPS-SIPRIS analysis of APE1 endonuclease bound to DNA with an abasic site analog (pdb_id: 5dfi) [54]. APE1 incises an abasic site phosphodiester backbone in DNA. A proposed mechanism involves superfamily-conserved active site residues forming hydrogen bonds with abasic site phosphate group oxygen atoms [55]. **A**. A cluster of EEP superfamily residues (yellow sidechains; $\boldsymbol{p = 5.2{\times}10^{-6}}$) and a hydrogen bond network of exoIII-AP-endo family residues (red sidechains; $\boldsymbol{p = 1.6{\times}10^{-6}}$), both of which are centered on the abasic site (circled). Family residues aggregate between the catalytic core and a loop containing DNA-interacting basic residues (magenta sidechains); they insert into the DNA major groove to form a kink that engulfs the abasic-DNA strand and thus appear to form a substrate-specific "reaction chamber". Nitrosation of exoIII-AP-endo residue Cys310 results in dissociation of APE1 from DNA and relocation to the cytoplasm [56]; thus, the associated hydrogen-bond network may communicate the nitrosation signal to the DNA binding site. **B**. Close up of the APE1 active site. (For clarity not all residues are shown.) Hydrogen bond atoms use CPK coloring.

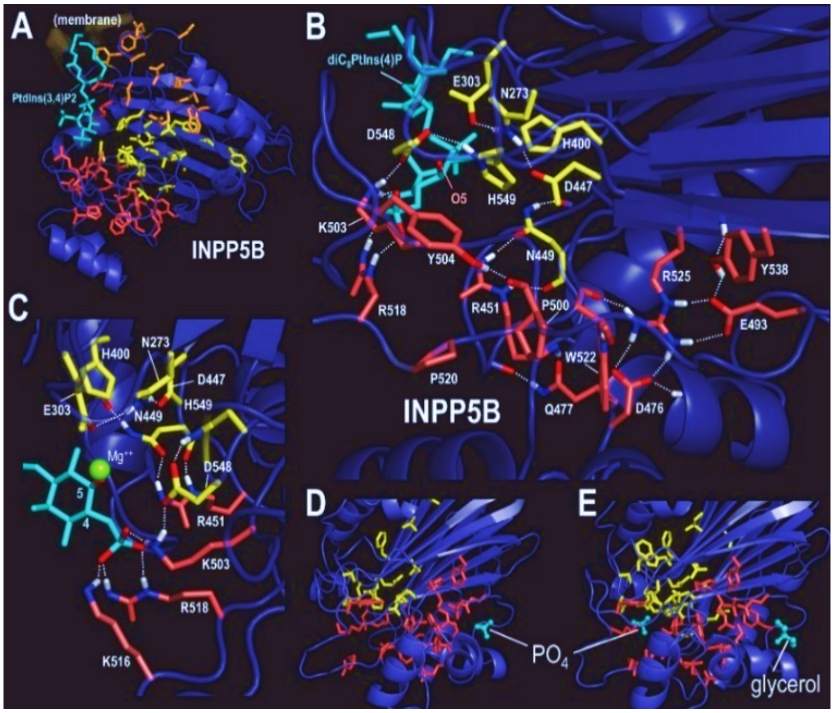
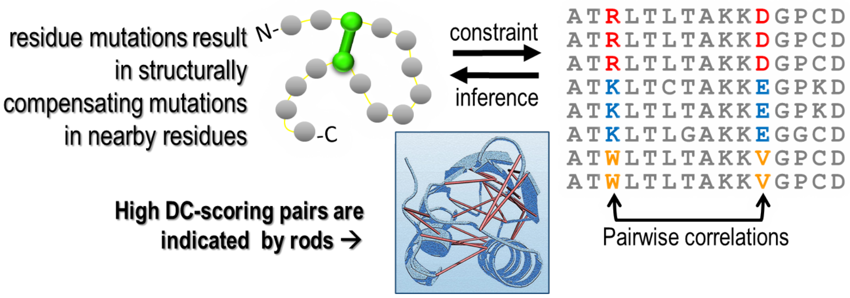


Figure 7.4. BPPS-SIPRIS analysis of the inositol polyphosphate 5-phosphatase INPP5B. <u>Color scheme</u>: EEP superfamily, INPP5 family and INPP5B subfamily residue sidechains: *yellow*, *red,* and *orange*, respectively; ligands, *cyan*; hydrogen bond atoms, *CPK coloring*. **A,B**. An INPP5 *<u>hydrogen bond network</u>* ($\boldsymbol{p = 1.1{\times}10^{-7}}$) that forms a secondary shell around the active site and is hypothesized to recognize inositol polyphosphates having phosphate groups attached at positions 4 and 5 of the inositol ring. This network is adjacent to a superfamily *<u>hydrogen bond network</u>* ($\boldsymbol{p = 3.2{\times}10^{-9}}$). **A**. INPP5B bound to the reaction product, phosphatidylinositol 3,4-bisphosphate (pdb_id: 4cml)[57]. SPIRIS *<u>clustering</u>* results: EEP, $\boldsymbol{p = 5.8{\times}10^{-13}}$; INPP5, $\boldsymbol{p = 3.9{\times}10^{-7}}$; INPP5B, $\boldsymbol{p = 0.0021}$. The INPP5B subfamily network lies between the proposed membrane interface [57] and the EEP catalytic core, suggesting a role in sequestering specific membrane-associated substrate from the lipid bilayer [58]. **B**. INPP5 hydrogen bond network within INPP5B (pdb_id: 3mtc). **C**. View focusing on the substrate 4-phosphate group. INPP5 enzymes cleave the 5-phosphate, but require for recognition the 4-phosphate, which directly interacts with three INPP5-family basic residues (K503, K516, and R518). **D**. In INPP5B (pdb_id: 5a7i [59]), INPP5-family residues most remote from the catalytic core are part of a cleft to which a phosphate is bound. This site may bind a molecule similar to the known substrate and may be allosterically linked to the active site via the network of INPP5 residues. **E**. The INPP5B-like OCRL protein with glycerol bound to a site analogous to that indicated in (D) (pdb_id:.4cmn [57]).

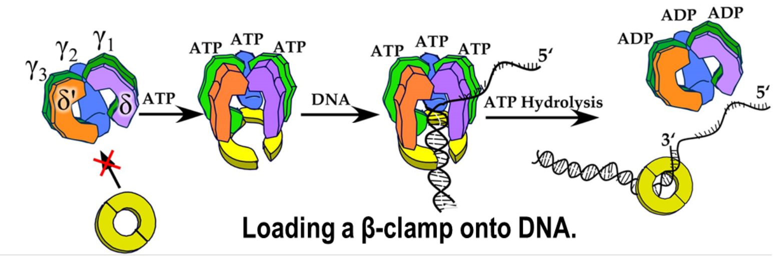


Direct coupling analysis. Predicting residue-to-residue structural contacts from a multiple alignment covariance matrix has been a topic of study for some time [60,61], the rationale being that residue mutations occurring at one position often result in compensatory mutations at other, structurally interacting positions. A problem with this straightforward approach is that residue positions may be correlated transitively: if position *i* interacts with position *j* and *j* with *k*, then residues at *i* and *k* may be correlated even though they fail to interact directly. A critical breakthrough came with **Direct Coupling Analysis (DCA)** [62-66], which distinguishes direct from indirect correlations [64,67,68]. DCA has been validated through numerous studies on globular [65,67,69] and membrane [62,63,70,71] proteins, and on protein internal repeats [72] and complexes [73-75]. **STARC (Statistical Tool for Analysis of Residue Couplings)** [76] estimates the statistical significance (as $S = -\log_{10}(p)$) of the correspondence between high scoring DC-pairs and 3D structural contacts; it can be used to evaluate DCA methods [76] and the potential functional relevance of alternative conformations and of homomeric and heteromeric interactions. Our DARC and SPARC programs (see below) incorporate both DCA and STARC. Prior to running STARC they perform DCA using pseudo-likelihood maximum entropy optimization [77], which outperformed [76] DCA methods based on sparse inverse covariance estimation [78] and on multivariate Gaussian modeling [79].

7. Multidimensional query-centric analysis of protein constraints (DARC)

DARC [80] (Deep Analysis of Residue Constraints) performs a multidimensional analysis that combines DCA, STARC, BPPS, and SIPRIS into a single, query-centric program for identifying and visualizing constraints as superfamily and functional-subgroup conserved residues, as family-specific, high DC-scoring residue pairs, and as correlations of these with each other and with structure. It does this by: **(1)** characterizing BPPS pattern residues and high DC-scoring residue pairs most distinctive of subgroup along the query sequence's lineage. **(2)** Visualizing pattern residues within representative aligned sequences. **(3)** Automatically generating PyMOL [81,82] session files showing the structural locations of constrained residues and of high DC-scoring pairs. **(4)** Determining how functional subgroup-specific residues and high DC-scoring pairs correlate with each other and with structure using SIPRIS. We illustrate DARC for bacterial DNA clamp loaders.

<u>*The bacterial DNA clamp loader AAA+ complex*</u> loads ring-shaped sliding β-clamps onto DNA to keep polymerase attached during replication; it contains one δ, three γ, and one δ' AAA+ subunits semi-circularly arranged in the order: δ-γ_1-γ_2-γ_3-δ'. Only γ is active, though both γ and δ' functionally influence an adjacent γ ATPase domain. Hence, γ and δ' share certain (γ/δ'-) constraints, while γ is subject to additional (γ-) constraints absent from δ'. In the presence of ATP, the clamp loader binds to and opens the β-clamp and, upon binding to DNA, ATP hydrolysis occurs, leading to closing of the β-clamp onto DNA [83-85].



Loading a β-clamp onto DNA.

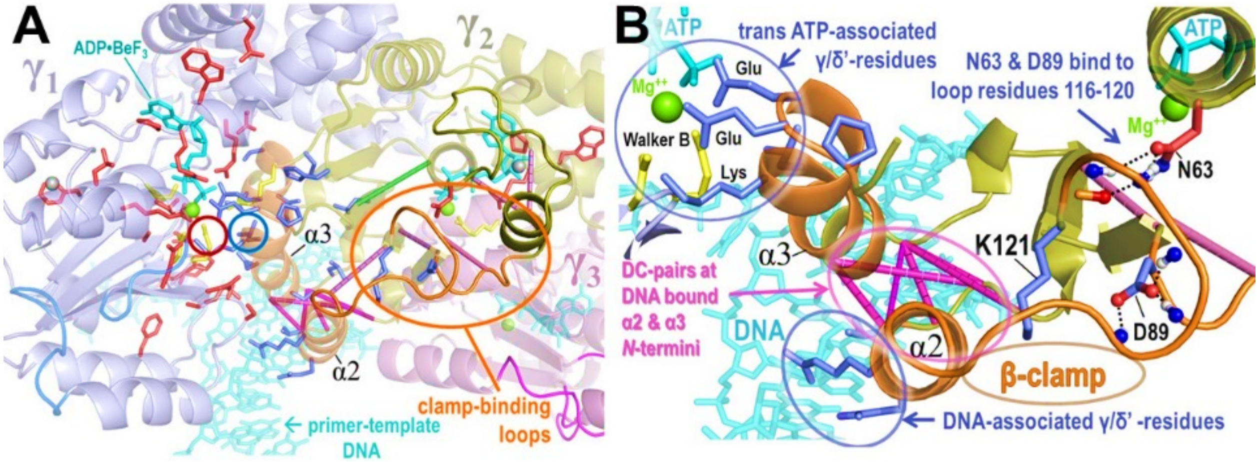

A
ADP•BeF3
γ1
γ2
α3
γ3
α2
clamp-binding loops
← primer-template DNA
B
ATP
trans ATP-associated γ/δ'-residues
N63 & D89 bind to loop residues 116-120
ATP
Mg++
Glu
Walker B
Glu
Lys
Mg++
N63
α3
K121
DC-pairs at DNA bound α2 & α3 N-termini
DNA
α2
D89
β-clamp
DNA-associated γ/δ' -residues

Fig. 9.1. Residue constraints within the *E. coli* DNA clamp loader complex bound to primer DNA and to an ATP analog (pdb_id: 3glf [86]). **A**. View of the γ- and γ/δ'-residues (*red* and *blue* sidechains, respectively) at the γ_1/γ_2 interface and of the top γ/δ'-DC-scoring pairs linking together (within γ_2) the DNA-binding α2 and α3 N-termini (*magenta* rods) and the β-clamp binding loops (*purple* rods). The γ-residues cluster around the catalytic base (*yellow* sidechain circled in red; $\boldsymbol{p=1.8\times10^{-9}}$). The γ/δ'-residues cluster around L140 (circled in blue; $\boldsymbol{p=6.7\times10^{-10}}$). **B**. Close up of constrained residues linking the ATP, DNA, and clamp binding sites. <u>*Color scheme*</u>: γ/δ'-residues, *blue*. γ-residues, *red*. Backbones of γ_1, γ_2 and γ_3, *blue*, *dark yellow*, and *pink*, respectively. β-clamp binding loops in γ_1, γ_2, and γ_3, *marine blue*, *orange*, and *bright pink*, respectively. Helices α2, α3, and α4 of γ_2, *orange*. Walker B catalytic residues in γ_1, *yellow*. ATP analog ADP•BeF_3 and DNA, *cyan*; Zn^{++}, *gray*; Mg^{++}, *green*.

The following functionally-congruent features interconnect the ATP, DNA, and clamp binding sites [80]: **(1)** Centered on the catalytic base near the γ-phosphate group of ATP are (SIPRIS-identified) γ- and γ/δ'-residue clusters on opposite sides of each γ-to-γ/δ' interface (**Fig. 9.1A**); the α3 helix C-terminus harbors γ/δ’-residues that interact with the adjacent ATP binding site (**Fig. 9.1B**). **(2)** The N-termini of the α2 and α3 helices bind DNA (due to their partial positive charges), harbor γ/δ’-residues interacting with DNA, and are entwined by six top γ/δ’-DC-scoring pairs (i.e., computed using a subgroup alignment of γ and δ’ sequences). And **(3)** the clamp-binding loop attaches to the α2 helix C-terminus, harbors two top γ/δ’-DC-scoring pairs, and forms hydrogen bonds with γ- and γ/δ’-residues. The positively charged sidechain of the γ/δ’-residue, K121, could interact with the α2 C-terminal negative dipole. Since DNA interacts with the α2 N-terminal positive dipole, α2 may act as a leveraging rod between the clamp binding site and DNA. Likewise, α3 may act as a rod linking the ATP-binding site to DNA. Together, these features appear to constitute an allosteric network coupling DNA binding to ATP hydrolysis & clamp loading [80]. The corresponding contrast alignments are shown in **Fig. 19.1** of Appendix 7: Using a MSA of 463,471 AAA+ proteins and the *E. coli* δ’ clamp loader subunit as the query, BPPS identified those residues that most distinguish both γ and δ’ from other AAA+ proteins (**Fig. 19.1C**). Using the γ/δ’ sub-MSA and the *E. coli* γ subunit as a query, BPPS identified those residues that most distinguish γ from δ’ (**Fig. 19.1E**). DARC saves a BPPS checkpoint file that can later be used to initiate a deeper analysis by expanding subtrees within the query’s lineage.

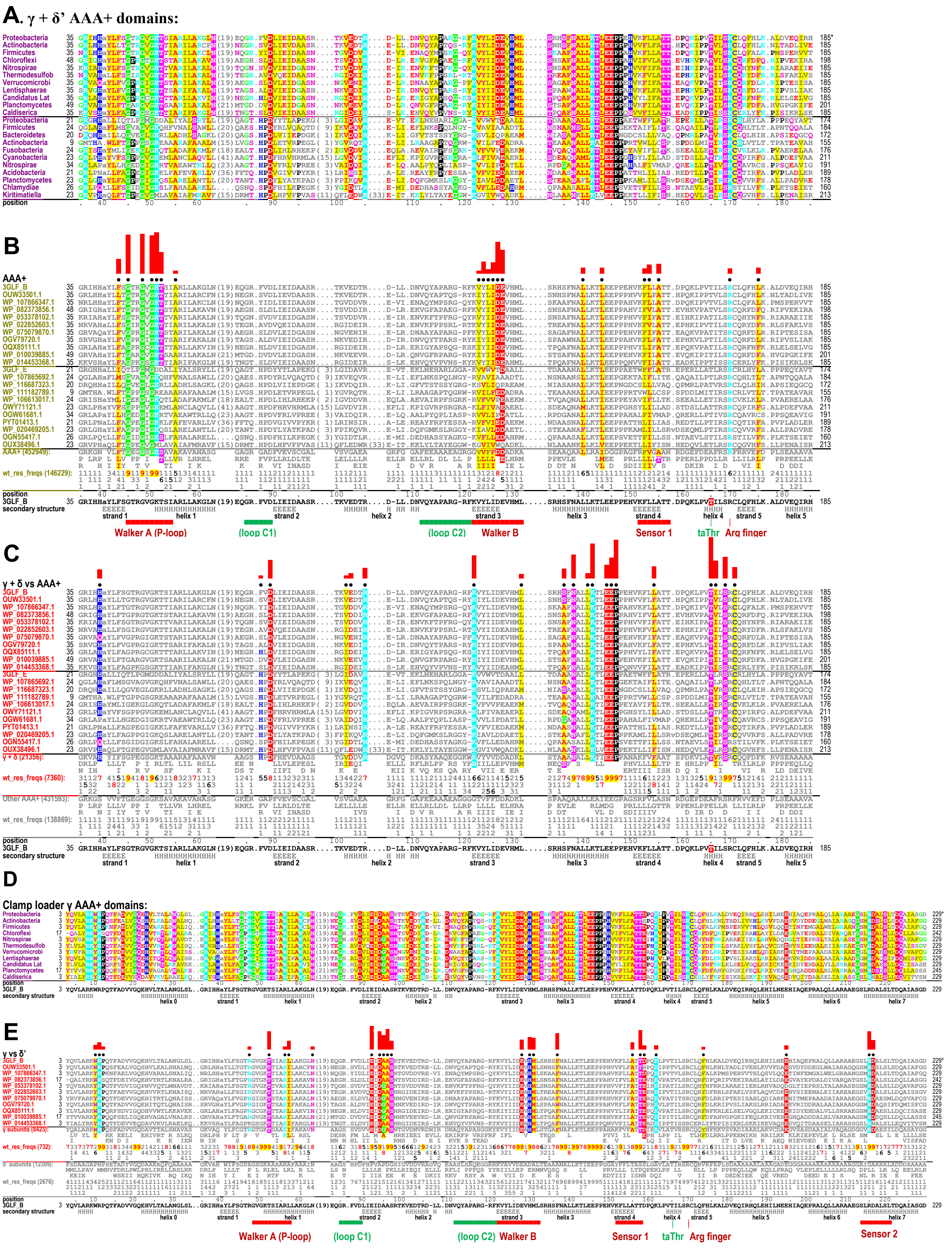


Figure 9.2. DARC-generated alignments highlighting all residues conserved in γ and δ’ clamp loader proteins and residues distinctive of the AAA+ superfamily, of the γ + δ’ subgroup, and of γ but not δ’. These are shown using five versions of the

same representative set of γ proteins (in panles a-e) and of δ’ proteins (in panels a to c). Residues are highlighted to indicate amino acid biochemical properties based on the following color code: red font with yellow highlight, non-polar (AVILMWFY) ; blue font with yellow highlight, cysteine (C); red, acidic (DE); cyan, basic (KR); magenta, polar (STNQ); green, glycine (G); blue, histidine (H); black, proline (P). Non-conserved positions in panels (a) and (d) and non-pattern residues in panels (b), (c) and (e) are shown in gray font. The leftmost columns in panels (b), (c) and (e) give the NCBI sequence identifiers; these are colored the same as the residue sidechains in Figure 2 of the paper. a. Alignment highlighting all γ + δ’ conserved residues. Those sequences above the line within the alignment correspond to representative γ proteins from the (distinct) phyla denoted in the leftmost column; the first sequence corresponds to the E. coli γ subunit (pdb_id: 3glfB). Those sequences below the line correspond to representative δ’ proteins from distinct phyla, the first sequence of which corrsponds to the E. coli δ’ subunit (pdb_id: 3glfE), which was used as the DARC query. The positions listed at the bottom correspond to the E. coli γ subunit. b. BPPS contrast alignment showing the same sequences as in panel (a), but highlighting only those residues most distinctive of the AAA+ superfamily. The heights of the red bars above each highlighted column estimate the selective pressure imposed on pattern residues at that position using a semi-logarithmic scale. Directly below the aligned sequences, the characteristic AAA+ residues at each position are shown and, directly below these, corresponding frequencies are given in integer tenths. A ‘7’, for example, indicates that the corresponding residue occurs in 70-80% of the 452,949 AAA+ sequences in the alignment. Below this is shown the residue positions and sequence of the E. coli γ subunit (with the Thr 165 residue that was mutated to Val highlighted in red), and shown below these are predicted secondary structure elements (symbol: H, helix; E, strand), helix and strand designations, and AAA+ structural motifs (red font) and putative clamp binding loops C1 and C2 (green font). Secondary structure assignments were calculated for the E. coli γ subunit using DSSP [87]. c. BPPS contrast using the same format as in panel (b) to highlight those residues that most distinguish γ and δ subunits from other AAA+ proteins. d. DARC-generated alignment highlighting all residues conserved in γ. e. DARC-generated alignment highlighting residues distinguishing γ subunits from δ’ subunits. A few of these are conserved in other catalytically active AAA+ ATPases; see panel (b).

8. Constraint analysis of empirical and MD simulated structures (SPARC)

SPARC (Search Procedures for Analysis of Residue Constraints) [88], like BPPS, runs in various modes. These modes fall into two categories: (1) the computation of DCA and STARC *S*-scores for one or more structures; and (2) Finding interactions within a time series set of MD simulated structures of a specific protein or protein complex. The following table lists the modes for each of these categories:

mode	Description
	Perform DCA and compute STARC S-scores
rank	Rank various protein structures based on *S*-scores.
hetmer	Compute *S*-scores for interacting heteromeric subunits (need fasta seqs with NCBI taxids)
simul	Compute *S*-scores for a time series set of MD simulated structures.
	Find interactions within a time series set of MD simulated structures
dist	Residue-to-residue or residue-to-ligand heavy atom contact distances
correl	Show contingency tables for correlated residue interactions and report interactions that form and dissociate during the simulation
sc2sc	Residue sidechain hydrogen bond interactions
sc2bb	Residue sidechain to backbone hydrogen bond interactions
sc2sb	Run both sc2sc and sc2bb modes; also reports CH-pi and aromatic-aromatic interactions
bb2bb	Find changes in backbone c=o to backbone n-h distances

SPARC rank mode. In the rank mode, SPARC takes as input an MSA in cma format and the path to a directory containing coordinate files for aligned sequences of known structure; it makes associations based on sequence/structural identifiers. Specifically, structural file names must use the syntax <pdbid>_H.pdb where <pdbid> is the lowercase version of the NCBI pdb identifier (e.g., '1abc' for pdb identifier 1ABC_A). The results for a rank search of death domain structures are shown in **Table 10.1**. **Fig. 10.1** shows the locations of the highest scoring direct couplings at homodimeric interfaces.

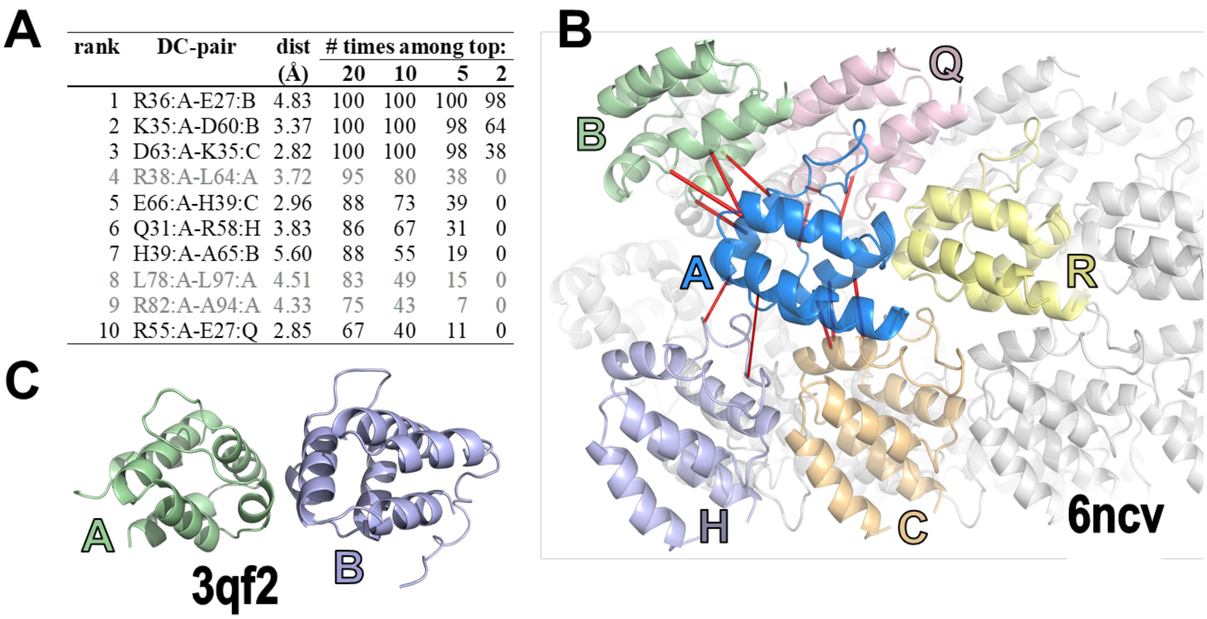
Table 10.1. SPARC ranking of pyrin-related death domain structures by *S*-score. Eighteen proteins of known structure were identified among 3,572 pyrin domain aligned sequences, five of which are shown. Search parameters: $r = 4.0$ Å; $m = 5$. See **Table 10.2** for parameter definitions. A colon between two chain designations (e.g., A:C) indicates that *S* was computed using, for each residue pair, the shorter of the internal versus the homodimeric 3D distances (e.g., the A-to-A versus the A-to-C residue distances).

pdbid	chain(s)	*S*	*L*	*D*	*X*	*d*	*F*	Δ*S*	resolution	method	Description
6ncv	A:C	42.0	1977	111	141	60	1.9	12.9	3.7 Å	cryo-EM	NLRP6 filament
6ncv	A:B	41.5	1977	111	141	60	1.9	12.4	3.7 Å	cryo-EM	NLRP6 filament
6ncv	A:H	37.1	1977	107	141	55	1.9	8.0	3.7 Å	cryo-EM	NLRP6 filament
6ncv	A:Q	36.6	1977	107	154	57	2.0	7.6	3.7 Å	cryo-EM	NLRP6 filament
2n1f	A:B	35.2	1977	98	141	53	1.8	11.8	4.0 Å	cryo-EM	ASC filament
6ncv	A	29.1	1977	92	141	46	1.9		3.7 Å	cryo-EM	NLRP6 filament
2n1f	A:G	28.2	1977	87	181	50	2.3	4.8	4.0 Å	cryo-EM	ASC filament
6ncv	A:R	27.7	1977	102	141	47	1.9	-1.4	3.7 Å	cryo-EM	NLRP6 filament
2n1f	A	23.4	1977	80	181	44	2.3		4.0 Å	cryo-EM	ASC filament
2n1f	A:H	22.7	1977	88	141	41	1.8	-0.7	4.0 Å	cryo-EM	ASC filament
4ewi	A	19.4	2045	103	151	42	1.9		2.28 Å	X-ray	NLRP4
3qf2	A	18.9	2045	101	187	45	2.4		1.7 Å	X-ray	NALP3
4ewi	A:B	18.7	2045	107	179	45	2.3	-0.6	2.28 Å	X-ray	NLRP4
5h7n	A:B	18.2	2042	100	185	44	2.4	0.2	1.85 Å	X-ray	NLRP12
5h7n	A	18.0	2042	97	185	43	2.4		1.85 Å	X-ray	NLRP12

3qf2	A:B	18.0	2045	105	187	45	2.4	-0.9	1.7 Å	X-ray	NALP3

Table 10.2. List of variables defined for STARC *S*-scores.

Symbol	Definition
L	Total number of MSA column pairs used
r	Maximum 3D distance used to define contacting residue pairs (default: 4 Å)
D	Number of contacting pairs, i.e. distinguished elements
X	Optimum cut point (as defined by STARC) for partitioning an array of length L
d	Number of left-distinguished elements, i.e. contacting pairs to the left of the cut point X (inclusive)
m	Minimum sequence separation between residue pairs in query protein of known structure
ℓ	The length of the input MSA
F	$F = X \div \ell$ indicates how spread-out the value of X is relative to the MSA length
S	-$\log_{10}$ P, where P corresponds to the estimated probability after correcting for multiple tests
ΔS	Change in the value of S upon the inclusion of interactions between homomeric subunit interface(s)



rank	DC-pair	dist	# times among top:			
		(Å)	20	10	5	2
1	R36:A-E27:B	4.83	100	100	100	98
2	K35:A-D60:B	3.37	100	100	98	64
3	D63:A-K35:C	2.82	100	100	98	38
4	R38:A-L64:A	3.72	95	80	38	0
5	E66:A-H39:C	2.96	88	73	39	0
6	Q31:A-R58:H	3.83	86	67	31	0
7	H39:A-A65:B	5.60	88	55	19	0
8	L78:A-L97:A	4.51	83	49	15	0
9	R82:A-A94:A	4.33	75	43	7	0
10	R55:A-E27:Q	2.85	67	40	11	0

Figure 10.1. SPARC *rank* analysis of pyrin-relate death domain (DD) proteins. **A**. Table of the 10 top residue pairs for the cryo-EM structure of the NLRP6 PYD filament (pdb_id: 6ncv [89]) based on sub-sampling of aligned pyrin-related sequences. SPARC robustly ranked residue pairs based on the number of times they were among the top DC-scoring (i.e., having the top average product corrected Frobenius norms) for 100 input MSA sub-samplings with replacement. Each sub-MSA sampled consisted of 500 sequences randomly drawn from among the 3,572 sequences in the input MSA. Seven of the 10 highest ranked pairs (those shown in black font) correspond to interactions that include contacts between adjacent death domains—suggesting that these contacts are functionally important. **B**. Image of the NLRP6 PYD filament cryo-EM structure. The 12 pairs that interact in trans, among the 30 highest ranked pairs, are indicated by red rods. Subunits adjacent to the A chain are colored, whereas non-adjacent subunits are shown in light gray. **C**. Image of the NALP3 PYD crystal structure (pdb_id: 3qf2 [90]). For this structure, SPARC computes a negative value for Δ*S* suggesting that the interaction lacks biological relevance and thus may be a crystallographic artifact.

Analyzing heteromeric complexes. SPARC's hetmer mode characterizes constraints associated with a heteromeric complex for which the user provides an MSA for each of two components. From each MSA, SPARC creates a subalignment consisting of the one sequence from each species that is most closely related to the corresponding 'gold standard' component within a given Cryo-EM or crystal structure of the complex. Finally, it computes S-scores for trans interactions between the two subunits, as illustrated in **Fig. 10.2**. Of course, some components may be absent for some species. To help identify such cases, SPARC outputs, for each component, a histogram of the pairwise scores between each of the candidates and its gold standard sequence—with scores for true orthologs tending to follow a unimodal distribution that is approximately normal. A published study [88] reports strong DC-constraints at several heterodimeric enzyme interfaces, indicative of selective pressures maintaining the residue couplings.

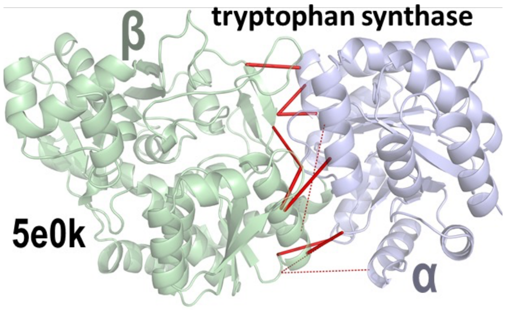


Fig. 10.2. The highest DC-scoring residue pairs across the α and β subunits of tryptophan synthase (pdb_id: 5e0k [1]) occur at the interface; $S = \log_{10}(p) = 21.1$. Solid and dashed red lines correspond to DC-pairs that are separated by ≤ 10 Å and > 10 Å, respectively.

Dynamic analysis of constraints. SPARC aids the interpretation of residue constraints by characterizing them dynamically rather than merely within static structures. Unlike standard analyses of molecular dynamics (MD) simulation data [91-97], this can reveal residue interactions and allosteric couplings associated with otherwise overlooked constraints. We illustrate this process for the *bacterial enhancer binding protein (bEBP) NtrC1* from *Aquifex aeolicus*: bEBPs activate transcription by remodeling RNA polymerase (RNAP) containing the sigma factor σ^{54} [98,99]. Simulations based on various NtrC1 ATP/ADP-bound states reveal alternative interactions involving residues distinctive of bEBP ATPases. For instance, when ATP is bound at the A:B interface and ADP at other interfaces, the AAA+ R-finger R299 interacts with the γ-phosphate of ATP (**Fig. 10.3, *right***), whereas at the adjacent F:A interface, R299 forms a salt bridge with the bEBP-residue D295—thereby sequestering it away from ADP, which then may be more easily expelled (**Fig. 10.3, *middle***). In the same state, the bEBP-residue R293 in subunit B forms a salt bridge with the catalytic base E239 of subunit A, pulling it away from the γ-phosphate of ATP to presumably inhibit ATP hydrolysis (**Fig. 10.4, left**). However, this salt bridge is disrupted when the nucleotide binding site of subunit F is vacant (**Fig. 10.4, right**). Together, these interactions may prevent ATP hydrolysis at the A:B interface until ADP is expelled from the F:A interface, perhaps thereby facilitating sequential hydrolysis around the hexameric ring. These interactions, which are absent in the corresponding crystal structure [100], were stably maintained during these 1 μs simulations.

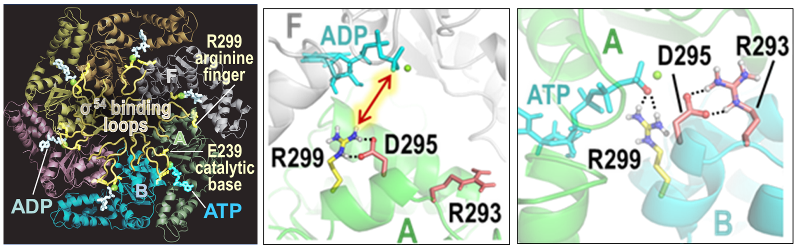


Fig. 10.3. (*left*) R299 and E239 within the bEBP hexamer bound to 1ATP + 5 ADP. (*middle*) The subunit A D295-R299 interaction may aid ADP release from subunit F. (*right*) The subunit B D295-R293 interaction frees the R-finger R299 to interact in trans with the γ-phosphate of ATP bound to subunit A.

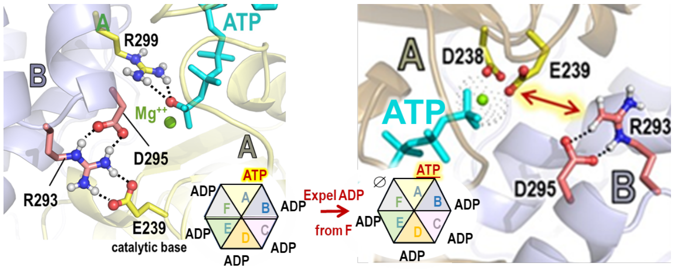


Fig. 10.4. MD conformations of R293 & E239 at the A:B interface in the 1ATP+5ADP (*left*) & 1ATP+4ADP+apo (*right*) states. The R293-E239 salt bridge may prevent ATP hydrolysis until ADP is expelled from the adjacent F subunit.

Correlated motion & allosteric coupling. SPARC can identify correlated interactions among constrained residues, as we illustrate here for NtrC1, which forms either a heptameric or a hexameric complex. Partial ATP occupancy causes the heptameric closed ring of NtrC1 to rearrange into a hexameric split ring that drives both ATP hydrolysis and the interaction of RNAP with σ^{54} [100]. During an MD simulation of the heptamer (pdb_id: 3m0e [101]) formation and dissociation of R293-D239-trans and R293-D295-cis interactions are highly correlated: When one salt bridge is formed, the other often dissociates, which may help mediate the heptamer-to-hexamer transition. An MD simulation of the hexamer (pdb_id: 4ly6 [100]) suggests that allosteric coupling of an R201:E246 cis-to trans switch to ATP-hydrolysis (**Fig. 10.5**) may mediate RNAP-σ^{54} remodeling.

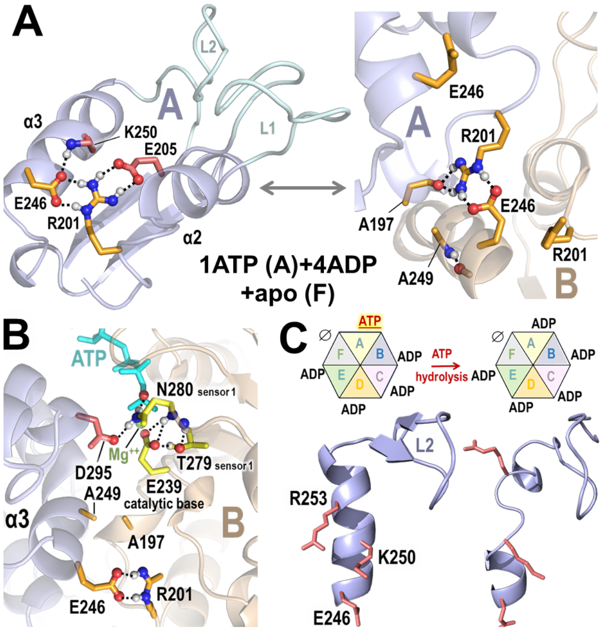


Fig. 10.5. Potential allosteric coupling of ATP hydrolysis to movement of the α2 and α3 helices, which are linked to the σ^{54}-binding L1 and L2 loops—as observed during a 1 μs MD simulation of the bEBP hexamer. **A**. A E246-R201 cis-to-trans switch associated with bEBP-residues, K250 and E205 (red), and with the two highest DC-scoring pairs, E246-R201 and A197-A249 (orange). **B**. The trans state is associated with formation of catalytically favorable interactions among sensor 1 residues, the catalytic base, and the ATP-γ-phosphate. **C**. Restructuring upon ATP hydrolysis of the α3 helix and of the σ^{54}-binding L2 loop.

9. Auxiliary programs

Analyses based on our major programs, namely MAPGAPS, BPPS, SIPRIS, DARC, SPARC, and eCOMPASS, are facilitated by the auxiliary programs listed in **Table 11.1**. Several of these are described in more detail below. The usage statement for each program provides further information.

Table 11.1. List of auxiliary programs that augment our main analyses.

program	description
addphylum	Adds taxonomic information to a file of fasta sequences; requires NCBI nr and taxdump files.
bpps2vsi	Creates a vsi (visualize structural interactions) file from files generated by bpps (see vsi2pml below).
cdd2mgs	Converts CDD hiMSAs (available at ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/hiMSA) into MAPGAPS format.
convert_msa	Converts an MSA from fasta to cma format and vice versa; also converts an MSA from cma into rtf format.
edit_cma	Performs editing operations on a cma-formatted MSA file.
edit_hpt	Performs editing operations on a bpps hyperpartition (*.hpt) file.
get_pdb	Automatically retrieves and adds hydrogen atoms to pdb coordinate files corresponding to pdb_identifiers within a cma-formated MSA or a fasta sequence file.
getPDB	Retrieves pdb coordinate files given a list of pdb_identifiers.
matchcma	Reports matches to user specified pattern residues within a set of bpps subgroup MSAs
purgemsa	Reduces sequence redundancy given a cma-formatted MSA or concatenated MSAs.
tree2hpt	Converts a tree in Newick format into a hyperpartition that can serve as input for bpps in H mode.
twkcma	(tweakcma) Performs a wide variety of operations on a cma-formatted MSA.
tweakPDB	Performs a wide variety of operations on a pdb structural coordinate file.
vsi2pml	Creates PyMOL scripts (*.pml) or session (*.pse) files from a vsi file.

Retrieving structural coordinates. Structural coordinate files can be retrieved from the PDB using the program **get_pdb**, which performs the following steps: It first retrieves NCBI **pdbaa** identifiers from sequences within either an input MSA or fasta file. Next, it retrieves corresponding structural coordinate files from the PDB and then runs the reduce program to add modeled hydrogen atoms. (Adding modeled hydrogen atoms aids identification of geometrically-accurate hydrogen bond interactions.) The names of final coordinate files take the form <pdb_id>_H.pdb, where <pdb_id> corresponds to the pdb identifier in lower case. This process requires both the fasta file **pdbaa**, which needs to be formatted using the command “makeblastdb -in $FASTADIR/pdbaa -input_type fasta -dbtype prot -parse_seqids”, and the programs **blastdbcmd** and **makeblastdb**, which may be downloaded from the NCBI via anonymous ftp at ftp.ncbi.nlm.nih.gov. The **blastdbcmd** program must be on your path with the environmental variable $FASTADIR set to the path to the directory containing the pdbaa file. You also need to put the script **batch_download.sh** (available at: www.rcsb.org) on your path and to set the environmental variable 'REDUCE_PRGM' to the **reduce** program, which is available at https://github.com/rlabduke/reduce.

Examining and editing structural coordinate files. The **tweakPDB** program takes as input a structural coordinate file in pdb format or (for a few options) a file listing the paths to multiple coordinate files. Its many different options allow the user to extract information or to make changes to the input file. For example, such information includes: covalent and hydrogen bond angles and distances; residue interactions with other residues or with substrate; residue buried surface areas; and amino acid sequences corresponding to protein chains. Possible changes to the input file include: converting MSE (selenomethionine) residues to MET or HIE, HID, HIP to HIS; identifying; eliminating water molecules; renaming chains; truncating chains; and renumbering residues in a chain. Such modifications may be necessary to complete an analysis.

Counting matching residues. The bpps program partitions an MSA into subgroups, which are stored as a concatenated series of sub-alignments within the output file **<prefix>_new.mma** (where the input alignment is named **<prefix>.mma**). Given this output file, the program **matchcma** computes the percentage of matching residues for each subgroup given a user-provided residue pattern corresponding to column positions in the MSA. For example, the command:

```
matchcma matchcma AAA+_new.mma W4,YF8,D193,P10,E70,R77
```

with AAA+_new.mma being a set of subalignments of AAA+ domains, returns the following percentages of conservation among 64 subgroups of AAA+ proteins.

```
ID:    PATTERN:     W4    YF8   D193    P10    E70    R77  Total_sq AvePercent
 1        RfcA:   90.2   92.0   97.2   97.7   97.4   98.6    1726.8 (95.5)+
 7        RfcC:   87.3   94.2   96.6   94.8   96.4   98.6    1168.2 (94.6)+
 6        RfcB:   91.2   92.6   94.6   94.6   97.2   97.4     927.5 (94.6)+
 8        RfcD:   88.2   87.7   95.2   92.6   98.4   99.2    1274.2 (93.6)+
 5        RfcS:   93.5   92.6   95.0   96.2   85.5   87.9    3072.7 (91.8)+
             :      :      :      :      :      :      :         :      :
14         MCM:    0.2    3.9    0.5    0.0      .    0.0    8121.3 (0.8)-
15        DnaA:      .    0.6    3.0    0.0      .    0.5   23669.8 (0.7)-
39    Dynein1:    0.2    1.2    1.2    1.0      .      .    1846.8 (0.6)-
19         Clp:    0.0    0.6    0.2    0.8    0.1    1.1   72440.2 (0.5)-
24        MoxR:    0.1    0.9    0.0    0.0      .    0.0   57145.5 (0.2)-
21        ClpX:    0.0    0.4    0.0    0.0    0.1      .   34122.7 (0.1)-
       PATTERN:     W4    YF8   D193    P10    E70    R77  Total_sq AvePercent
```

In this case, the pattern matches various replication factor C (Rfc) AAA+ subunits.

Visualizing structural features. Often an analysis involves dozens or hundreds of protein structures. To avoid creating too much output, several of our programs create a *.vsi file, which stores the information required to visualize results using PyMOL. The **vsi2pml** program is used to obtain PyMOL scripts (i.e., *.pml files), which requires a separate coordinate file, or, if PyMOL is on your path, session files (i.e., *.pse), which include the coordinates. This allows the locations of and interactions among various categories of residues to be visualized using PyMOL, as illustrated in other sections of this tutorial.

10. Performing a complete sequence/structural analysis.

This section gives step-by-step instructions on how to perform a sequence/structural analysis using our approach. Details regarding the input, output and runtime options are described in the command line usage statement for each of our program.

Compiling our source code. The following items are used to compile our source code, which is available at https://www.igs.umaryland.edu/labs/neuwald/software/

JSON library (https://github.com/json-c); this is required for compilation.
CUDA (https://developer.nvidia.com/cuda-downloads) is required to run direct coupling analysis faster on a GPU; but this is optional.
OpenMP for multiprocessing (though any decent compiler should already include it).
Both bison and flex are required:
Download the src for e.g. bison from http://ftp.gnu.org/gnu/bison/
Download flex from https://github.com/westes/flex/releases

CMake (https://cmake.org/download/) is required for compilation. On many systems it should be already installed. To automatically obtain structural coordinate files and to model hydrogen atoms using our get_pdb program, two third-party programs need to be installed and on your path (see below).

Obtaining and annotating sequence data. The NCBI non-redundant (nr) and pdbaa fasta and taxdump files are available via anonymous ftp at ftp://ftp.ncbi.nih.gov/. You can access this site using a command line FTP program, login as "anonymous" and give your email address as your password. Once at this cite use the following commands to obtain these files:

```
cd blast/db/FASTA
get pdbaa.gz
get nr.gz
cd ../../../
cd pub/taxonomy/accession2taxid/
get pdb.accession2taxid.gz ${PROJECT_HOME}/molbio/ncbi_downloads/pdb.accession2taxid.gz
get prot.accession2taxid.gz ${PROJECT_HOME}/molbio/ncbi_downloads/prot.accession2taxid.gz
cd ../
get taxdump.tar.gz ${PROJECT_HOME}/molbio/ncbi_downloads/taxdump.tar.gz
```

Move these files to appropriate directories and use **gunzip** to decompress them. You should set the environmental variables $FASTADIR and $TAXDUMPDIR to these directories so that our programs can access these files. To taxonomically annotate the fasta files, you first must extract the taxonomy files using the commands:

```
tar xvf taxdump.tar
grep 'scientific name' names.dmp > scientific.dmp
```

and then use the following commands to taxonomically annotate sequences in the fasta files.

```
cat nr | addphylum stdin > nrtx
cat pdbaa | addphylum stdin -pdb > pdbaatx
```

To access curated hierarchical MSAs (hiMSAs), which are used as queries by our MAPGAPS program, again log onto the NCBI ftp site, go into the hiMSA directory (cd pub/mmdb/cdd/hiMSA/) and download the files you require.

You can speed up MAPGAPS searches by running many jobs concurrently over grid nodes, in which case you will need to first split the nrtx fasta file into subfiles using the command and then combine and merge the output (cma-formated) files using our tweakcma program with the -m option.

```
fasplit nrtx 250000 < nrtx
```

This will create many subfiles, designated as: nrtx.1, nrtx.2, nrtx.3, …, containing up to 250,000 sequences each. MAPGAPS searches these files to create input MSAs for analysis by out other programs.

Obtaining structural coordinate (pdb) files. This can be done automatically using our program **get_pdb**. The command 'get_pdb <infile>', where <infile> is a fasta file containing pdbaa (fasta) sequences, returns the corresponding structural coordinate files in pdb format and then adds modeled hydrogen atoms to the files. Two files will be created for each pdb identifier: the original pdb-formated file (e.g., pdb1abc.ent) and a smaller file that is used by our programs and that includes modeled hydrogen atoms. This requires that two third-party programs be on your path, namely: **batch_download.sh**, which retrieves the files, and **reduce**, which adds modeled hydrogen atoms. These are available, respectively, at:
https://www.rcsb.org/docs/programmatic-access/batch-downloads-with-shell-script
and https://github.com/rlabduke/reduce.
Our program **tweakPDB** can be used to further modify or analyze pdb files.

Obtaining an input MSA for analysis by our programs. If you have downloaded an hiMSA curated by the NCBI CDD group (see above), you can convert this into **MAPGAPS**-format using the **cdd2mgs** program. The output files can then be used to search for and multiply align sequences belonging to the superfamily modeled by the hiMSA. Alternatively, if no such hiMSA is available, you can use a cma-formated MSA (obtained using **GISMO** or another MSA program) to create a hiMSA using the following steps:

```
cp <infile>.cma <infile>.tpl
foreach file ($FASTADIR/nrtx.*)  mapgaps <infile> file
cat $FASTADIR/nrtx.*_map.seq > All
mapgaps <infile> All
mapgaps All_map                    // creates *mpa and other files
foreach file ($FASTADIR/nrtx.*)  mapgaps All_map file
mapgaps All_map $FASTADIR/pdbaatx
cat  $FASTADIR/pdbaatx_A.cma $FASTADIR/nrtx.*_A.cma > X.cma
tweakcma X -m                      // merges the concatenated cma files
tweakcma X.merged -mincol=0.80     // removes sequences fragments
tweakcma X.merged.match80 -U90     // reduces sequence redundancy to < 90% identity
\mv -f tweakcma X.merged.match80.purge90.cma Main.cma
twkcma Main -hsw                   // creates a Main.hsw file with sequence weights
```

The resultant Main.cma and Main.hsw files are used as input to the bpps and darc programs. If the input MSA is in mfasta format, you can use **convert_msa** (in fa2cma mode) to convert it into cma format. Currently under development is the LAPIS (Lots of Accurately-aligned Proteins Initiated from Scratch) program, which performs the above steps automatically starting from a fasta file of sequences belonging to a given protein superfamily. To heuristically reduce sequence redundancy the faster tweakcma -

cdhit=<int> option may be used. The relative quality of MSAs may be assessed using **eCOMPASS**. An input MSA may be modified using the **tweakcma** or **edit_cma** programs.

Partition the MSA into subgroups based on conserved patterns. Running **BPPS 1** globally partitions an MSA into divergent subgroups based on subgroup-specific patterns. To evaluate consistency among runs use the **BPPS E** mode. To expand the alignment within each partition use **BPPS 2** and **BPPS 3** modes, which will create a hiMSA of the superfamily. If the input MSA is huge, BPPS can take considerable time, in which case **DARC** may be used to create the lineage within an implicit hierarchy for a specific query sequence. DARC also runs a **STARC** (DCA) analysis and creates both a **SIPRIS** (*.sprs) input file and a *vsi file. **SIPRIS** identifies and computes the significance of pattern residue structural clusters. The **vsi2pml** program takes as input a *.vsi file to create PyMOL files to visualize the structural locations of residue constraints. For both BPPS and DARC a *.terms output file to assess the fasta define terms and taxonomic information for each partition. (Can also create a terms file using **tweakcma** with the -terms option.)

Defining and refining models of protein superfamily residue constraints. BPPS creates both a *.hpt file, describing the partitions and corresponding patterns, and a *.sma file, containing a seed sequence for each partition. BPPS also has an option (-heatmap) that creates asset of heat map, one for each subgroup showing the degree to which that subgroup's pattern is conserved both within that subgroup and withing other subgroups. Such heat maps may indicate that the hierarchy should be further edited (possibly into a non-hierarchical 'hyperpartition') to model more accurately the constraints imposed on the superfamily. (The heatmap PyMOL files can reveal whether certain subgroups share constrained residues in a non-hierarchical manner.) Such editing may be performed using either our **edit_hpt** program and/or a text editor, such as vim. Likewise, the corresponding *.sma file may be edited. Given such *.hpt and *.sma files, the **BPPS H** mode may be used not only to refine a partition, but—given an input MSA of the same length as the original MSA—to also update the overall bpps model as addition sequences become available. This allows models of protein superfamilies to be maintained and refined over time without having to start all over again from scratch. Of course, high quality hiMSA models may also be used as MAPGAPS queries.

Characterizing residue constraints dynamically. In a cellular context, of course, proteins typically undergo conformational changes that may not be observed in available structural models. To explore the dynamic structural properties associated with BPPS-defined residues and high DC-scoring residue pairs one may perform molecular dynamics simulations on proteins of interest. We typically set up such simulations using **CHARMM-GUI** (https://www.charmm-gui.org/) and run the simulation using **OpenMM** (https://openmm.org/). The **SPARC** program can be used to search the set of simulated (or also empirical) structures for interactions of interest.

Interpret results in the light of published research. One should read the literature on proteins under investigation to help interpret results.

Appendix 1: Gibbs Sampler for Multi-alignment Optimization (GISMO)

GISMO.

Notation and Definitions. The following notation is used for vectors $\mathbf{v}=\left(v_1,...,v_n\right)^T$ and $\mathbf{w}=\left(w_1,...,w_n\right)^T$: $\left|\mathbf{v}\right|=\left|v_1\right|+...+\left|v_n\right|$, $\mathbf{v}+\mathbf{w}=\left(v_1+w_1,...,v_n+w_n\right)^T$, $\mathbf{v}/\mathbf{w}=\left(v_1/w_1,...,v_n/w_n\right)^T$, $\mathbf{v}^{\mathbf{w}}=v_1^{w_1}...v_n^{w_n}$, and $\Gamma\left(\mathbf{v}\right)=\Gamma\left(v_1\right)...\Gamma\left(v_n\right)$. Given *K* proteins, their sequences are defined by $\mathbf{R}=\left(R_1^T,...,R_K^T\right)^T$ where each vector $R_k=\left(r_{k,1},...,r_{k,n_k}\right)$ corresponds to the *k*-th sequence, n_k is the *k*-th sequence's length and the $r_{k,i}$ corresponds to the *i*-th residue in that sequence. $\mathbf{h}(\ )$ defines a counting function where, for example, $\mathbf{h}\left(R_k\right)$ returns a length 20 vector of the counts for the residue types in R_k.

A block-based alignment of the input sequences is defined by w columns. The set of variables defining the sequence positions for column *j* is defined by $A_j=\left\{a_{1,j},...,a_{K,j}\right\}$. We define $A_{j[-k]}\equiv A_j-\left\{a_{k,j}\right\}$ to denote the set A_j without $a_{k,j}$. An alignment is defined by the matrix $\mathbf{A}=\left(A_1,...,A_w\right)^T$ and $\left\{\mathbf{A}\right\}\equiv\left\{a_{k,j}:k=1,..,K,j=1,...,w\right\}$ denotes the set of residues indices for the alignment variable **A**. We represent the collection of residues indexed by elements in a set C as $\mathbf{R}_C$. For instance, $\mathbf{R}_{\{\mathbf{A}\}}=\left\{a_{k,j}:k=1,...,K;j=1,...w\right\}$ represents the set of residues in the alignment defined by **A**.

GISMO Statistical Model. The residue frequencies observed for column *c* are modeled as a multinomial distribution with parameters $\boldsymbol{\theta}_c=\left(\theta_{1,c},...,\theta_{20,c}\right)^T$ where $\sum_{i=1}^{20}\theta_{i,c}=1$and $\theta_{i,c}>0$ for all *i*. That is, the vector $\boldsymbol{\Theta}=\left(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_w\right)$ defines a product multinomial model corresponding to the full alignment. The vector $\boldsymbol{\theta}_0$ corresponds to a background amino acid residue distribution. Hence, the complete-data likelihood function is given by

$$\pi\left(\mathbf{R}\middle|\boldsymbol{\theta}_0,\boldsymbol{\Theta},\mathbf{A}\right)\propto\boldsymbol{\theta}_0^{\mathbf{h}(\mathbf{R})}\prod_{j=1}^{w}\left(\frac{\boldsymbol{\theta}_j}{\boldsymbol{\theta}_0}\right)^{\mathbf{h}\left(\mathbf{R}_{\{A_j\}}\right)}$$

where it is assumed that $\Theta\sim D\left(\mathrm{B}\right)$ and $\boldsymbol{\theta}_0\sim D\left(\boldsymbol{\alpha}\right)$ (where D denotes the Dirichlet distribution), and where $\mathrm{B}=\left(\beta_1,...,\beta_w\right)$ specifies the Dirichlet distribution parameters (commonly interpreted as numbers of pseudocounts) at each column position *j*, and $\boldsymbol{\alpha}$ specifies the parameters for the background distribution. (Recall that the alignment is specified by the matrix $\mathbf{A}=\left(A_1,...,A_w\right)=\left(a_{k,j}\right)_{K\times w}$ where $a_{k,j}$ indicates the position of the *j*-th column, which is assumed to be present in all of the sequences.) The likelihood of **A** with the $\boldsymbol{\theta}$'s integrated out is

$$\pi\left(\mathbf{R}\middle|\mathbf{A}\right)\propto\Gamma\left(\mathbf{h}\left(\mathbf{R}_{\{\mathbf{A}\}^c}\right)+\boldsymbol{\alpha}\right)\cdot\prod_{j=1}^{w}\Gamma\left\{\mathbf{h}\left(\mathbf{R}_{\{\mathbf{A}_j\}}\right)+\boldsymbol{\beta}_j\right\}. \tag{1}$$

The conditional predictive probability distribution of this conserved region occurring at position *i* in sequence *k* is given by

$$\pi\left(a_k = i \middle| \mathbf{A}_{[-k]}, \mathbf{R}\right) \propto \prod_{j=1}^{w} \left(\frac{\hat{\boldsymbol{\theta}}_j}{\hat{\boldsymbol{\theta}}_0}\right)^{\mathbf{h}\left(r_{k,a_{k,j}}\right)}$$

where the $\hat{\boldsymbol{\theta}}$ are the posterior means of the $\boldsymbol{\theta}$, given the observed sequence data **R** and the current alignment $\mathbf{A}_{[-k]}$. This statistical model serves as the foundation for the HMM [10] used in later stages of sampling.

Dirichlet mixture priors. In order to capture the fact that certain biochemically or structurally similar amino acid residues are more likely to occur together we have incorporated Dirichlet Mixture priors [102,103], as refined by [104]. In order to speed up sampling, GISMO uses a 20 component mixture in the first (competitive) phase of sample, inasmuch as the goal is to merely obtain a reasonable starting alignment without overtraining the evolving HMM. After this initial phase GISMO applies a 58-component mixture.

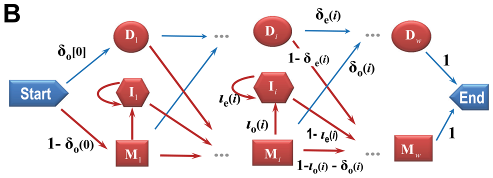
Down weighting for sequence redundancy. Sequences are down weighted for redundancy using the following procedure. For each sequence *k* a non-integer weight is computed using the method of Henikoff and Henikoff [105] as:

$$wt(k) = \sum_{j=1}^{w} \left(Nt_j \cdot Nr_{k,j}\right)^{-1}$$

where Nt_j is the number of residue types at each position *j* and where $Nr_{k,j} = \left|\left\{ r_{a_{x,j}} \middle| 1 \le x \le K \wedge r_{a_{x,j}} = r_{a_{k,j}} \right\}\right|$ is the number of sequences with the same residue at position *j* as for sequence *k*. The rational for this formulation is that if a sequence matches lots of sequences at most positions, then it should receive a lower weight than a sequence that matches few sequences at most positions. These weights are then normalized and integerized as:

$$Wt(k) = \left\lceil 100 \cdot Wt(k) \div wt_{\max} \right\rceil$$

where $wt_{\max}$ corresponds to the maximum non-integer sequence weight. Because these weights depend upon the evolving alignment, they are updated after each sampling cycle.



Inferred HMM transition probabilities. We model the transition probabilities for the HMM (shown on the right) using a generalization of our previous formulation [10] as follows. The probability matrix for transitions from column *j* states in the HMM is:

$$\begin{array}{cccc}
 & \mathrm{M}_{j+1} & \mathrm{I}_j & \mathrm{D}_{j+1} \\
\mathrm{M}_j & 1-\iota_o[j]-\delta_o[j] & \iota_o[j] & \delta_o[j] \\
\mathrm{I}_j & 1-\iota_e[j] & \iota_e[j] & 0 \\
\mathrm{D}_j & 1-\delta_e[j] & 0 & \delta_e[j]
\end{array}$$

where $1 \le j \le w$ and where M, I, and D denote match, insertion and deletion states, respectively. The probability matrix for transitions out of the start state is:

$$\begin{array}{ccc}
 & \mathrm{M}_1 & \mathrm{D}_1 \\
\text{Start} & 1-\delta_o[0] & \delta_o[0]
\end{array}$$

Transitions into M and I states emit a residue as specified by the Θ of our statistical model.

*Inference of transition probabilities***.** For a given alignment ***A***, each sequence S_k is associated with a "path" through the HMM indicating its alignment against the model Θ. We denote the collection of these paths by Λ and the total number of HMM transitions of type M→M, M→I , … , D→D at position *j* by

$$N_{mm}[j],\ N_{mi}[j],\ N_{md}[j],\ N_{im}[j],\ N_{ii}[j],\ N_{dm}[j] \text{ and } N_{dd}[j].$$

Ignoring the indexing variable j for clarity, the likelihood of the transition probability parameters at each position is

$$h(\Lambda \mid \iota,\delta) = (1-\iota_o-\delta_o)^{N_{mm}}\,\iota_o^{N_{mi}}\,\delta_o^{N_{md}}\,(1-\iota_e)^{N_{im}}\,\iota_e^{N_{ii}}\,(1-\delta_e)^{N_{dm}}\,\delta_e^{N_{dd}}.$$

with independent prior distributions

$$(\iota_o,\delta_o,1-\iota_o-\delta_o) \sim D(n_{mi},n_{md},n_{mm}),\ \iota_e \sim \mathrm{Beta}(n_{ii},n_{im}), \text{ and } \delta_e \sim \mathrm{Beta}(n_{dd},n_{dm}),$$

where $n_{mi}, n_{md}, n_{mm}, n_{ii}, n_{im}, n_{dd}, n_{dm}$ are corresponding prior pseudo counts. The corresponding maximum a posteriori probability (MAP) estimates for the transition probabilities at each position j are computed from these observed and prior counts. These define the position specific gap penalties. The joint posterior distribution for the alignment and transition probability parameters is

$$g(\mathbf{A},\Lambda,\vec{\boldsymbol{\iota}},\vec{\boldsymbol{\delta}}) \propto P(\mathbf{R} \mid \mathbf{A},\Lambda) \times h(\Lambda \mid \vec{\boldsymbol{\iota}},\vec{\boldsymbol{\delta}}) \times P(\vec{\boldsymbol{\iota}},\vec{\boldsymbol{\delta}}),$$

where $P(\mathbf{R} \mid \mathbf{A},\Lambda)$ is a generalization of Equation (1), and where $\vec{\boldsymbol{\iota}}$ and $\vec{\boldsymbol{\delta}}$ are length w vectors representing the column-specific transition probabilities with prior probability:

$$P(\vec{\boldsymbol{\iota}},\vec{\boldsymbol{\delta}}) = \left[D(n_{mi},n_{md},n_{mm}) \times \mathrm{Beta}(n_{ii},n_{im}) \times \mathrm{Beta}(n_{dd},n_{dm})\right]^{w}.$$

Given the alignment and thus the paths Λ, we have the conditional posterior distribution

$$p(\vec{\boldsymbol{\iota}},\vec{\boldsymbol{\delta}} \mid \mathbf{A},\Lambda) \propto \prod_{j=1}^{w}\Big[\iota_o[j]^{N_{mi}[j]+n_{mi}-1}\cdot\delta_o[j]^{N_{md}[j]+n_{md}-1}\cdot(1-\iota_o-\delta_o[j])^{N_{mm}[j]+n_{mm}-1}\times$$
$$\iota_e[j]^{N_{ii}[j]+n_{ii}-1}(1-\iota_e[j])^{N_{im}[j]+n_{im}-1}\cdot\delta_e[j]^{N_{dd}[j]+n_{dd}-1}(1-\delta_e[j])^{N_{dm}[j]+n_{dm}-1}\Big]$$

Sampling on the distribution for each position j is done by drawing the random variables:

$$\delta_o[j] \sim \mathrm{Beta}(N_{md}[j]+n_{md},N_{mm}[j]+N_{mi}[j]+n_{mm}+n_{mi}),$$
$$\delta_e[j] \sim \mathrm{Beta}(N_{dd}[j]+n_{dd},N_{dm}[j]+n_{dm}),$$
$$\iota_o[j] = (1-\delta_o[j])\iota_o^*[j], \text{ where } \iota_o^*[j] \sim \mathrm{Beta}(N_{mi}[j]+n_{mi},N_{mm}[j]+n_{mm}),$$
$$\text{and } \iota_e[j] \sim \mathrm{Beta}(N_{ii}[j]+n_{ii},N_{im}[j]+n_{im}).$$

For computational efficiency, the ι and δ may be integrated out [10] to get

$$\begin{aligned} h(\Lambda) &= \iint h(\Lambda \mid \vec{\boldsymbol{\iota}},\vec{\boldsymbol{\delta}})P(\vec{\boldsymbol{\iota}},\vec{\boldsymbol{\delta}})\,d\vec{\boldsymbol{\iota}}\,d\vec{\boldsymbol{\delta}} \\ &= \prod_{j=1}^{w}\Bigg[\frac{\Gamma(N_{mi}[j]+n_{mi})\Gamma(N_{md}[j]+n_{md})\Gamma(N_{mm}[j]+n_{mm})\Gamma(n_{m\cdot})}{\Gamma(N_{m\cdot}[j]+n_{m\cdot})\Gamma(n_{mi})\Gamma(n_{md})\Gamma(n_{mm})} \\ &\quad\times\frac{\Gamma(N_{ii}[j]+n_{ii})\Gamma(N_{im}[j]+n_{im})\Gamma(n_{i\cdot})}{\Gamma(N_{im}[j]+N_{ii}[j]+n_{i\cdot})\Gamma(n_{ii})\Gamma(n_{im})} \\ &\quad\times\frac{\Gamma(N_{dd}[j]+n_{dd})\Gamma(N_{dm}[j]+n_{dm})\Gamma(n_{d\cdot})}{\Gamma(N_{dd}[j]+N_{dm}[j]+n_{d\cdot})\Gamma(n_{dd})\Gamma(n_{dm})}\Bigg]. \end{aligned}$$

This gives rise to a new posterior distribution $g(\mathbf{A},\Lambda) \propto P(\mathbf{R} \mid \mathbf{A},\Lambda) \times h(\Lambda)$, for which the transition probability parameters need not be fixed or updated and which allows the optimal indel penalties to be determined from the sequence data.

Sampling algorithm. GISMO's MCMC sampling algorithm explores the space of possible alignments by executing Markovian transitions between alignments. This involves sampling alternative alignments of either individual sequences or groups of sequences. In either case, such sampling is done as follows: First, the sequence or sequences are removed from the alignment and the posterior parameters of the HMM are recalculated based on the retained aligned sequences and the priors. Next, emission probabilities for the twenty amino acids at each position are sampled from the posterior emission probability distributions defined by the HMM parameters; note that these sampled probabilities define a sampled HMM. Finally, the previously removed sequences are optimally realigned to the sampled HMM. We explored sampling transition probabilities in the same way, but found little benefit of doing so; instead, the MAP estimates for transition probabilities are used. GISMO applies simulated annealing [22] to favor convergence on an optimal alignment in later stages of sampling. Sampling starts at a "temperature" of $T = 1$ (i.e., sample each transition directly proportional to its actual probability p) and ends at $T = 0$ (i.e., always take the highest probability transition); between these two extremes the temperature is dropped in $\Delta T = 0.1$ increments with sampling probabilities set to $p^{1/T}$. Sampling iteratively through all of the sequences continues until this fails to find a new highest probability state.

Appendix 2: Bayesian Partitioning with Pattern Selection (BPPS)

Bayesian Partitioning with Pattern Selection (BPPS). Given a typically very large multiple sequence alignment (MSA), denoted here as $\mathbf{X}$, BPPS applies Markov chain Monte Carlo (MCMC) sampling to articulate a superfamily into a set of hierarchically nested partitions corresponding to a tree. Each subtree *h*, which may consist of only a single node, when attached to the root corresponds to a family, and, in general, when attached to a parent node corresponds to a child subgroup. The sampler defines each subgroup, also denoted by *h*, based on residue patterns distinguishing subgroup members from sequences assigned to other nodes in the parent subtree. For instance, a simple pattern for subgroup *h* might consist of $\{\mathrm{V,I,L}\}$, $\{\mathrm{D,E}\}$, and $\{\mathrm{F,Y}\}$ at column positions 3, 10 and 23, respectively. BPPS favors assignment of those sequences to subtree *h* conserving a pattern that is not conserved in sequences assigned to other nodes in the parent subtree. Hence, BPPS favors assignment to each parent node those sequences conserving the parent node's pattern but lacking each of the descendent nodes' patterns. For a non-root node *n* in the hierarchy this process defines a 'contrast' alignment (as in **Fig. 6.2C,D**) divided into foreground and background sequences, corresponding respectively to the subtree rooted at *n* and to the rest of the subtree rooted at the parent of *n*. MCMC sampling is used to determine the number and arrangement of the nodes in the tree, the sequences belonging to each node, the pattern positions for each subgroup, and the conserved residues at each pattern position. The sampler favors convergence on a hierarchy where the pattern defining the partitioning for each node best distinguishes its foreground from its background. BPPS weights sequences, as in PSI-BLAST[105,106], to avoid modeling conserved patterns merely due to sequence redundancy.

For the ensuing discussion, we define the following: For vectors $\mathbf{v}=(v_1,...,v_\ell)^T$ and $\mathbf{w}=(w_1,...,w_\ell)^T$, $\mathbf{v}/\mathbf{w}=(v_1/w_1,...,v_\ell/w_\ell)^T$, $\mathbf{v}+\mathbf{w}=(v_1+w_1,...,v_\ell+w_\ell)^T$, $\log\mathbf{v}=(\log v_1,...,\log v_\ell)^T$, $|\mathbf{v}|$ is the sum over vector elements, and $\langle\mathbf{v},\mathbf{w}\rangle$ denotes the inner product of **v** and **w** and is equivalent, as applied here, to the dot product $\mathbf{v}\cdot\mathbf{w}=\sum_{i=1}^{\ell} v_i w_i$. Given an *N* node hierarchy $\mathbf{H}$, we define **S** as a vector of *N* disjoint sets, such that $\mathbf{S}_n$ contains the sequences assigned to node *n*, and $\mathbf{H}$ is defined as a vector of tri-partitions of node indices $1\le n\le N$, such that $\mathbf{H}_h\equiv\langle H_h^+,H_h^-,H_h^o\rangle$ specifies subtree *h*'s foreground, background and "non-participating" nodes, respectively. We define node *n* = 1 as the root and *h* = 1 as the superfamily tree (i.e., $H_1^+=\{n\,|\,1<n\le N\}$), for which the background set ($H_1^-=\{0\}$) consists of a single node (*n* = 0) for unrelated sequences (denoted as $\mathbf{S}_0$). The remaining $\mathbf{H}_h$ are configured hierarchically starting from the root, such that: H_h^+ specifies the nodes in subtree *h*; H_h^- specifies the nodes in the tree rooted at the parent node of *h* but absent from H_h^+; and H_h^o specifies nodes in neither H_h^+ nor H_h^-. To ensure that $\mathbf{H}$ corresponds to a tree, we require that each H_h^+, other than H_1^+, is a proper subset of only one other $H_{h'}^+$ (i.e., $\forall h:h>1\rightarrow\exists!h':H_h^+\subset H_{h'}^+$) and that H_h^- consist of nodes in $H_{h'}^+$ that are not in H_h^+ $\left(\text{i.e.}, H_h^-=H_{h'}^+-H_h^+\right)$.

BPPS defines a prior on **H** that depends only on *N* and positive parameter *v* and that assumes a maximum number of nodes $N_{\max}$, so that

$$p(\mathbf{H}) = \frac{p(N)}{a_N}, \text{ where } p(N) = \begin{cases} \nu^{N-1} \cdot (1-\nu)/(1-\nu^{N_{\max}}) & \text{if} \quad \nu \neq 1 \\ 1/N_{\max} & \text{if} \quad \nu = 1 \end{cases},$$

and where a_N, the number of unlabeled, unordered rooted trees with N nodes, is defined recursively as:

$$a_N = \begin{cases} 1 & \text{if} \quad N = 1 \\ \displaystyle\sum_{\substack{j_1 + 2j_2 + \ldots + (N-1)j_{N-1} \\ = N-1}} \prod_{k=1}^{N-1} \binom{a_k + j_k - 1}{j_k} & \text{if} \quad N > 1 \end{cases}$$

with j_k being the number of subtrees with k nodes [107]. Computation suggests that the growth of a_N is $O(2.96^N)$. By default, $N_{\max} = 500$ and $\nu = 1$ so that $p(\mathbf{H}) = (a_N N_{\max})^{-1}$, which corresponds to a uniform prior where every size tree (up to $N_{\max}$) is equally likely. Setting $\nu > 1$ or $\nu < 1$ favors hierarchies with more or fewer nodes, respectively. Note, however, that adding nodes when unjustified by the data is disfavored regardless of $p(\mathbf{H})$ due to the nature of our Bayesian formulation.

Let σ_n be the prior for a sequence assigned to node n. Given N, the prior for $\mathbf{S}$ is then given by $p(\mathbf{S}) = \prod_{n=0}^{N} \sigma_n^{|S_n|}$. By default, we choose a prior for the rejected sequence node of $\sigma_0 = 0.5$ and for other nodes $\sigma_n = (1-\sigma_0) \cdot N^{-1}$ uniformly.

Given $\mathbf{H}$ and $\mathbf{S}$, foreground pattern residue sets are denoted by $\mathbf{A}$, where $A_{h,c} \neq \varnothing$ for each pattern column position c in subgroup h, and $A_{h,c} = \varnothing$ for non-pattern positions. The pattern residue sets $\mathbf{A}_h$ are constrained by a foreground consensus sequence for subgroup h, denoted as $\mathbf{y}_h$, and by the requirement that the residues in each set be functionally similar, as were defined by a detailed analysis of amino acid Dirichlet mixture components[104]. For example, if a tryptophan residue occurs at position c of the consensus for H_h^+, then $y_{h,c} = \mathrm{W}$ and

$$\mathrm{A(W)} \equiv \{\{\mathrm{W}\}, \{\mathrm{W,F}\}, \{\mathrm{W,Y}\}, \{\mathrm{W,F,Y}\}\} \text{ and } A_{h,c} \in \mathrm{A(W)} \cup \{\varnothing\},$$

where $\mathrm{A}(r)$ denotes the allowed pattern residue sets for consensus residue r. We define prior probabilities for $\mathbf{A}$ as $p(\mathbf{A}) = \prod_{h=1}^{N} \prod_{c=1}^{C} \rho_{A_{h,c}}$ (product categorical distributions), where, when $A_{h,c} = \varnothing$, $\rho_{A_{h,c}} = q_{\varnothing}$ (0.999 by default) and otherwise $\rho_{A_{h,c}} = \kappa \cdot \frac{q^{|A_{h,c}|}}{\|A_{h,c}\|}$. Here, κ is a constant chosen so that the priors for pattern positions sum to $1 - q_{\varnothing}$; $0 < q < 1$ (and 0.5 by default) is a tuning parameter with smaller values yielding higher aggregate priors for the class of "functional" residue sets of smaller cardinality $|A_{h,c}|$; and $\|A_{h,c}\|$, defined as the number of possible residue sets of cardinality $|A_{h,c}|$, functions to distribute prior probabilities uniformly among these sets. For example, if $y_{h,c} = \mathrm{W}$ and $A_{h,c} = \{\mathrm{W,F}\}$, then $\|A_{h,c}\| = 2$ because, in this case, there are two possible sets of cardinality 2.

Since we are interested only in whether a residue in column *c* of a sequence *s* (i.e., $x_{s,c}$) is functional or non-functional, we introduce the variable $\boldsymbol{\chi}$, where $\boldsymbol{\chi}_{h,s,c} = (1,0)^{\mathrm{T}}$ and $\boldsymbol{\chi}_{h,s,c} = (0,1)^{\mathrm{T}}$ imply that, for subtree *h*, $x_{s,c}$ corresponds to a 'functional' and a 'non-functional' pseudo-residue, respectively. Since $A_{h,c} = \varnothing$ at non-pattern columns, corresponding residues are all non-functional. Gaps are also treated as non-functional.

Given **H**, **S** and **A**, let $\boldsymbol{\theta}_{h,c}$ be the 2-dimensional vector specifying the observed functional and non-functional pseudo-residue background frequencies for column *c* of subtree *h*, and let $\boldsymbol{\Theta}$ be the matrix of all $\boldsymbol{\theta}_{h,c}$. Let $\boldsymbol{\theta}_{h,c}^{(\alpha_h)} \equiv (1-\alpha_h)\boldsymbol{\theta}_{h,c} + \alpha_h (1,0)^{\mathrm{T}}$ model the foreground composition where $1-\boldsymbol{\alpha}$ specifies the fraction of background 'contamination' at pattern positions in the foreground. The prior probability density for α_h is defined by a beta distribution

$$p(\alpha_h) = \frac{\Gamma(a_{h,0}+b_{h,0})}{\Gamma(a_{h,0})\Gamma(b_{h,0})} \alpha_h^{a_{h,0}-1} (1-\alpha_h)^{b_{h,0}-1},$$

where $a_{h,0}$ and $b_{h,0}$ are functional and non-functional pseudo-counts, respectively, and where by default $a_{h,0} = b_{h,0} = 1$. The prior probability density for $\boldsymbol{\theta}_{h,c}$ is defined by a product Beta distribution:

$$p(\boldsymbol{\Theta}) = \prod_{h=1}^{N} \prod_{c=1}^{C} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta_{h,c,1}^{a-1} \theta_{h,c,2}^{b-1},$$

where *C* is the number of columns in the MSA, and $a = b = 1$ by default.

Conditional on **H**, $\boldsymbol{\chi}$ and **S**, let $\xi_{h,c}$ denote the inferred number of functional pseudo-residues in column *c* of subtree *h* that are *not* due to background contamination. Then, conditional on α_h and $\boldsymbol{\theta}_{h,c}$,

$$\xi_{h.c} \,|\, \alpha_h, \boldsymbol{\theta}_h \sim \mathrm{Binom}\left(Nf_{h,c}, \frac{\alpha_h}{\alpha_h + (1-\alpha_h)\theta_{h,c,1}} \right),$$

where $\theta_{h,c,1}$ and $Nf_{h,c}$ are, respectively, the background frequency and the total foreground number (with background contamination included) of the pattern-matching pseudo-residues in column *c* for subtree *h*. Conditional on **H**, $\boldsymbol{\chi}$, **S**, $\boldsymbol{\xi}_h \equiv (\xi_{h,1},\ldots,\xi_{h,C})^{\mathrm{T}}$, and $Nf_{h.} = \sum_c Nf_{hc}$, the posterior distribution of α_h is

$$\left[\alpha_h \,|\, \boldsymbol{\xi}_h\right] \propto \alpha_h^{|\boldsymbol{\xi}_h|+a_{h,0}-1} (1-\alpha_h)^{Nf_{h.}-|\boldsymbol{\xi}_h|+b_{h,0}-1} \sim \mathrm{Beta}\left(|\boldsymbol{\xi}_h| + a_{h,0}, Nf_h - |\boldsymbol{\xi}_h| + b_{h,0}\right).$$

The conditional distribution for $\boldsymbol{\theta}_{h,c}$ is:

$$\boldsymbol{\theta}_{h,c} \,|\, \mathbf{H},\mathbf{S},\boldsymbol{\xi}_h,\boldsymbol{\chi}_h \sim \mathrm{Beta}\left(\sum_{n \in H_h^-} \sum_{s \in S_n} \chi_{h,s,c,1} + Nf_{h,c} - \xi_{h,c} + \psi_1, \sum_{n \in H_h^- \cup H_h^+} \sum_{s \in S_n} \chi_{h,s,c,2} + \psi_2 \right),$$

where $\boldsymbol{\psi} \equiv (a,b)^{\mathrm{T}}$ specifies pseudo-counts with $\boldsymbol{\psi} = (1,1)^{\mathrm{T}}$ by default.

The sampler infers $\mathbf{H}, \mathbf{S}, \mathbf{A}, \boldsymbol{\alpha}$, and $\boldsymbol{\Theta}$ from $\mathbf{X}$, which defines an MSA. Given these variables, the logarithm of the joint probability distribution[23] is defined as:

$$\log P\left(\mathbf{X},\mathbf{H},\mathbf{S},\mathbf{A},\boldsymbol{\alpha},\boldsymbol{\Theta}\right)=\log P\left(\mathbf{X}\middle|\mathbf{H},\mathbf{S},\mathbf{A},\boldsymbol{\alpha},\boldsymbol{\Theta}\right)+\log p\left(\mathbf{H}\right)+\log p\left(\mathbf{S}\right)$$
$$+\log p\left(\mathbf{A}\right)+\log p\left(\boldsymbol{\alpha}\right)+\log p\left(\boldsymbol{\Theta}\right) \qquad (1)$$

where, assuming statistical independence among subtrees (but see below),

$$\log P\left(\mathbf{X}\middle|\mathbf{H},\mathbf{S},\mathbf{A},\boldsymbol{\alpha},\boldsymbol{\Theta}\right)=\sum_{h=1}^{N}\left(\sum_{n\in H_h^+\cup H_h^-}\sum_{s\in S_n}\sum_{c=1}^{C}\left\langle\log\boldsymbol{\theta}_{h,c},\boldsymbol{\chi}_{h,s,c}\right\rangle+\sum_{n\in H_h^+}\sum_{s\in S_n}\sum_{c=1}^{C}\boldsymbol{I}_{A_{h,c}}\left\langle\log\frac{\boldsymbol{\theta}_{h,c}^{(\alpha_h)}}{\boldsymbol{\theta}_{h,c}},\boldsymbol{\chi}_{h,s,c}\right\rangle\right) \qquad (2)$$

and where $\boldsymbol{I}_{A_{h,c}}\equiv\begin{cases}0, & \text{if}\quad A_{h,c}=\varnothing\\ 1, & \text{if}\quad A_{h,c}\neq\varnothing\end{cases}$. Note that for non-pattern positions $\boldsymbol{\theta}_{h,c}=\left(0,1\right)^{\mathrm{T}}$ and $\chi_{h,s,c}=\left(0,1\right)^{\mathrm{T}}$ so that $\left\langle\log\boldsymbol{\theta}_{h,c},\boldsymbol{\chi}_{h,s,c}\right\rangle=\log 0\cdot 0+\log 1\cdot 1=0$; hence, these positions contribute nothing to the posterior probability. Note too that the first summed term in Equation (2) is always negative, whereas if node *h* is assigned a pattern that is conserved in the foreground but not the background, then the second summed term will be net positive for node *h*, thereby probabilistically favoring that model.

Because the configuration of each subtree constrains to some degree the possible configurations of other subtrees above or below it in the hierarchy, our independence assumption is invalid. These constraints reduce the probabilities for some states to zero, so that the probabilities assigned to the remaining, reachable states will sum to less than 1, and our formulation is therefore conservative (i.e., computed probabilities are smaller than they should be). This also occurs due to other imposed constraints, such as placing an upper bound on the number of pattern positions or on the depth of the hierarchy, or requiring that a minimum number of sequences be assigned to each node. Nevertheless, in searching for an optimum, Equation 1 is valid as an objective function, its use here.

BPPS sampling strategies. Conditioned on fixed **H**, BPPS samples over **S** and **A** by iteratively applying the following. For each sequence *s*, let *n* be its assigned node and remove *s* from S_n. Then, sample *s* to a new node n' with probability proportional to $P\left(\mathbf{X},\mathbf{H},\mathbf{S},\mathbf{A},\boldsymbol{\alpha},\boldsymbol{\Theta}\middle|s\in S_{n'}\right)$ after having updated $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}$. Likewise, for each column position *c* in each subtree *h*, remove the pattern set $A_{h,c}$ and sample in a new pattern set $A'_{h,c}\in\mathrm{A}(\mathrm{y}_{h,c})\cup\left\{\varnothing\right\}$ with probability proportional to $P\left(\mathbf{X},\mathbf{H},\mathbf{S},\mathbf{A},\boldsymbol{\alpha},\boldsymbol{\Theta}\middle|A_{h,c}=A'_{h,c}\right)$ after having updated $\boldsymbol{\chi}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}$. However, if the number of pattern positions for a given subtree *h* is greater than a specified maximum $C_{\max}$ (25 by default), reduce the number down to $C_{\max}$ by removing the lowest probability pattern positions.

The BPPS sampler is initialized by setting $N=1$ with $\mathbf{H}_1=\left\langle H_h^+=\left\{1\right\},H_h^-=\left\{0\right\},H_h^o=\varnothing\right\rangle$, and assigning all sequences to the root node ($n=1$) with $C_{\max}$ pattern positions for subtree $h=1$ and with background pseudo-residue frequencies at each position (the $\boldsymbol{\theta}_{0,c}$) derived from the overall residue frequencies for the entire MSA. At this stage, sampling over **S** merely involves iteratively assigning sequences either to the foreground $\left(S_1\right)$ or to the background $\left(S_0\right)$, where the background represents unrelated sequences inadvertently included in the alignment. Sampling over **A** generally tweaks pattern assignments slightly due to removal of unrelated sequences. This provides a good starting point to speed up convergence with essentially no risk of getting trapped in a suboptimal state. Convergence is defined by a cycle of sampling over **S** and **A** that fails to improve upon the best configuration found thus far, as defined by the log-probability (Equation 1). BPPS saves the best configuration for the final output.

After convergence with N = 1 nodes, a child node may be added to the root node, as follows. First, some of the sequences assigned to the root are reassigned to the child node by selecting a subset of sequences that are more similar to each other than they are to the remaining sequences. Selections are based on similarity to the query, when one is designated, and on similarity to an arbitrary sequence otherwise. Next, BPPS samples for a few cycles over **S** and **A**, as described above, to search for a configuration that improves upon the previous hierarchy based on Equation 1. If the hierarchy fails to improve and a query has not been provided, several other candidate queries may be selected in turn until either an improved state is found or until a prespecified number of attempts are tried. If the hierarchy is improved, BPPS further enlarges and rearranges the evolving hierarchy **H** by adding more leaf nodes and by deleting, inserting or moving nodes using this same basic strategy. To avoid excessively complex hierarchies, BPPS requires that each leaf node contain a minimum number of sequences (50 by default); those that do not are pruned. After convergence, the sampler applies simulated annealing[22] to 'drop into' a more nearly optimal configuration.

When DARC applies BPPS, it focuses on the query's lineage within the superfamily hierarchy by first defining the query's family based on residues that most distinguish family members from other superfamily members. Next, DARC seeks to recursively define, in a similar manner, the query's subfamily, and other subgroups further down the query's lineage to a prespecified maximum depth.

Appendix 3: Direct Coupling Analysis (DCA)

Direct Coupling Analysis (DCA). DARC performs DCA using the algorithm implemented in the CCMpred program version 0.3.2 (https://travis-ci.org/soedinglab/CCMpred)[77], which is essentially identical to the plmDCA[66] and GREMLIN[108] algorithms and which we modified to output DCA scores in PSICOV format. Our description here follows closely the one given for CCMpred[77]. The rationale behind DCA is that, over evolutionary time, mutations at a given residue position are compensated for by mutations at interacting positions to thereby maintain structural integrity. DCA works by avoiding the confounding effect of indirect correlations due, for example, to two residues both interacting with a third residue, but not with each other. DARC uses the sub-MSA defined by BPPS to compute the highest scoring directly coupled residue pairs (DC-pairs).

The CCMpred algorithm eliminates indirect interactions from an interaction network by inferring a generative model of the MSA based on a Markov Random Field (MRF). We again represent the input MSA as an *R* row × *C* column matrix **X**, where element $x_{s,c}$ corresponds to the residue in row (i.e., sequence) *s* and column *c*. The columns correspond to vertices of the MRF with single-residue emission potentials $\varepsilon_c\left(r\right)$ for amino acid residue $r\in\left\{1,\ldots,20\right\}$ in column *c*; covariation between columns corresponds to edges of the MRF with pairwise emission potentials $\varepsilon_{c,d}\left(r_c,r_d\right)$ for residues r_c and r_d in columns *c* and *d*, respectively. In theory, one could optimize the parameters of the MRF given the MSA using as the objective function the probability:

$$P\left(\boldsymbol{\varepsilon}\middle|\mathbf{X}\right)=\frac{1}{Z}\prod_{s=1}^{R}\prod_{c=1}^{C}\left[\exp\left(\varepsilon_c\left(x_{s,c}\right)+\sum_{\substack{d=1\\d\neq c}}^{C}\varepsilon_{c,d}\left(x_{s,c},x_{s,d}\right)\right)\right]$$

where *Z* is a normalization constant to ensure that the sum over all sequences equals 1. However, because computing $P\left(\boldsymbol{\varepsilon}\middle|\mathbf{X}\right)$ is intractable for a non-trivial MSA, the following pseudo-log-likelihood is used instead as the objective function:

$$plL\left(\boldsymbol{\varepsilon}\middle|\mathbf{X}\right)=\sum_{s=1}^{R}\sum_{c=1}^{C}\left[\varepsilon_c\left(x_{s,c}\right)+\sum_{\substack{d=1\\d\neq c}}^{C}\varepsilon_{c,d}\left(x_{s,c},x_{s,d}\right)-\log Z_{s.c}\right]$$

where $Z_{s,c}=\sum_{r=1}^{20}\exp\left[\varepsilon_c\left(r\right)+\sum_{\substack{d=1\\d\neq c}}^{C}\varepsilon_{c,d}\left(r,x_{s,c}\right)\right]$.

Because computation of the normalization constants $Z_{s,c}$ involve summing over only *C* terms, these are much faster to compute than *Z* for $P\left(\boldsymbol{\varepsilon}\middle|\mathbf{X}\right)$. The gradient (the vector of partial derivatives) of this pseudo-log-likelihood is given by:

$$\frac{\partial plL\left(\boldsymbol{\varepsilon}\middle|\mathbf{X}\right)}{\partial\varepsilon_{c,d}\left(r,r'\right)}=\sum_{s=1}^{R}\left\{\delta_{x_{s,d},r'}\left[\delta_{x_{s,c},r}-\frac{1}{Z_{s,c}}\exp\left(\varepsilon_c\left(r\right)+\sum_{\substack{i=1\\i\neq c}}^{C}\varepsilon_{i,c}\left(r,x_{s,i}\right)\right)\right]\right\}$$

$$=\sum_{s=1}^{R}\left[\delta_{x_{s,d},r'}\right]\left[\delta_{x_{s,c},r}-p\left(x_{s,c}=r\middle|\left(x_{s,1},\ldots,x_{s,c-1},x_{s,c+1},\ldots,x_{s,C},V,E\right)\right)\right]$$

where $\delta_{x,y}$ is the Kronecker delta function.

In order to favor sparse solutions, we add an L_2 regularization term $R(\boldsymbol{\varepsilon})$ and maximize $plL(\boldsymbol{\varepsilon}|\mathbf{X}) - R(\boldsymbol{\varepsilon})$ using the nonlinear conjugate gradient method where

$$R(\boldsymbol{\varepsilon}) = \lambda_{\text{single}} \sum_{c=1}^{C} \left\| \varepsilon_c \right\|_2^2 + \lambda_{\text{pair}} \sum_{\substack{c,d=1 \\ d \neq c}}^{C} \left\| \varepsilon_{c,d} \right\|_2^2 ,$$

where the regularization coefficients are $\lambda_{\text{single}} = 1$; $\lambda_{\text{pair}} = 0.2 \times (L - 1)$ [108], and where $\left\| V_c \right\|_2^2$ and $\left\| E_{c,d} \right\|_2^2$ are the sum of squared residuals, which measure the discrepancy between the model and the data (with smaller values indicating a tighter fit of the model to the data).

After a successful optimization, the couplings between residue positions $K_{c,d}$ are ranked by the Frobenius norms of the edge potentials $\varepsilon_{c,d}$:

$$K_{c,d} = \sqrt{\sum_{r,r'=1}^{20} \varepsilon_{c,d}(r,r')^2}$$

Lastly, an Average Product Correction[109] is applied to arrive at the final score:

$$\zeta_{c,d} = K_{c,d} - \frac{K_{c,\cdot} K_{\cdot,d}}{K_{\cdot,\cdot}}$$

where "." denotes averaging over the corresponding row or column and $K_{\cdot,\cdot}$ is the average over all matrix elements.

Evaluating the robustness of DCA score rankings. To determine whether different input MSAs rank DC-pairs consistently, an auxiliary subsampling routine is included as an option in DARC. For the analysis here, this routine draws from the input MSA 1,000 samples of 1,000 sequences, from each of which DCA scores are computed. Between samplings, the previously sampled sequences are replaced prior to sampling the subsequent set. The percentage of times that each residue pair was among those with the top 20, 10, 5 or 2 DC-pairs is given in an output table. Those pairs consistently selected among the 20 top scores are also output.

Appendix 4: Initial Cluster Analysis (ICA)

Initial Cluster Analysis (ICA). To compute ${}_{CL}S_P$, ${}_{3D}S_{DC}$, ${}_{3D}S_P$, or ${}_{DC}S_P$ (denoted generically here as S) we apply Initial Cluster Analysis[110], a statistical approach to address the following question: Consider an array of 0s and 1s of length L and containing D 1s. Are some or all of the 1s significantly clustered near the start of the array, and, if so, how surprising is the most significant such clustering? To make this determination, ICA applies the Minimum Description Length (MDL) principle[111], an information theoretical regularization method for finding the best hypothesis for a given set of data.

The MDL principle defines a theory θ as a probability distribution P_θ over all possible sets of data and the description length of a data set E given a theory θ as $\mathrm{DL}(E|\theta) = -\log(P_\theta(E))$. A model M is a parameterized set of theories, and the description length of E given M is defined as $\mathrm{DL}(E|\mathrm{M}) = \min_{\theta\in\mathcal{M}} \mathrm{DL}(E|\theta)$. The MDL principle asserts that among multiple models to explain E, one should prefer the model M that minimizes $\mathrm{DL}(E|\mathrm{M}) + \mathrm{COMP}(\mathrm{M})$, where the description length or complexity $\mathrm{COMP}(\mathrm{M})$ is the log of the number of effectively independent theories M contains. For ICA, the MDL principle determines whether the hypothesis H_1 that the 1s cluster near the start of the sequence is better than the null hypothesis H_0 that the 1s and 0s occur randomly.

ICA treats H_1 as a single-parameter model, whose parameter x describes the location of a cut at a discrete point from 1 to $L-1$ along the array, thereby dividing it into an initial segment s_1 of length x, and a terminal segment s_2 of length $y = L - x$. If s_1 contains D_1 1s, and s_2 contains $D_2 = D - D_1$ 1s, assume that s_1 is generated by Bernoulli trials with maximum-likelihood probability $P_1 = D_1/x$ for a 1, and s_2 is generated by Bernoulli trials with probability $P_2 = D_2/y$ for a 1. Given a particular fixed value for x, the probability of E is $P_x(E) = P_1^{D1}(1-P_1)^{x-D_1} P_2^{D_2}(1-P_2)^{y-D_2}/Z$, where Z is a normalization constant taken over all length L sequences having D 1s. Hence the description length of E under H_1 is $\mathrm{DL}(S|\mathrm{H}_1) = -\log(\max_x P_x(E))$. ICA computes the complexity of H_1, as

$$\mathrm{COMP}(\mathrm{H}_1) \approx \log\left(\sqrt{D/\pi}\,\frac{L-1}{2}\right).$$

ICA treats H_0 as a model consisting of a single Bernoulli-trial theory for generating E, with the probability of a 1 taken as $P = D/L$, and of a 0 as $Q = 1 - P$. Hence, $\mathrm{DL}(E|\mathrm{H}_0) = -\log(P^D Q^{L-D})$, which is L times the entropy of the Bernoulli trial. Because H_0 contains only one theory, its complexity is zero. The MDL principle says that we should prefer H_1 to H_0 when $\mathrm{DL}(E|\mathrm{H}_1) + \mathrm{COMP}(\mathrm{H}_1) < \mathrm{DL}(E|\mathrm{H}_0)$. Treating each hypothesis as equally likely *a priori*, we may view the difference Δ between the two sides of this inequality as a log-odds ratio, and use the logistic function $\frac{e^\Delta}{1+e^\Delta}$ to convert this into a *p*-value (see p. 37 of [112]), from which $S = -\log_{10}(p)$ is defined.

Appendix 5: Structurally Interacting Pattern Residues' Inferred Significance (SIPRIS)

SIPRIS. SIPRIS relies on **Initial Cluster Analysis (ICA)**, as described in Appendix 4 and which addresses the following questions: Consider a string of 0s and 1s of length *L* and containing *D* 1s. Are some or all of the 1s significantly clustered near the start of the sequence, and, if so, how surprising is the most significant such clustering? Here we focus on the statistical and information theoretical bases of ICA as applied to BPPS-SIPRIS analyses.

BPPS-defined residue sets. Modes 2-3 of BPPS generate a hiMSA (**Fig. 6.1**). For each subgroup (i.e., subtree) *G* within a hierarchy, BPPS defines a corresponding set of "discriminating" residues that most distinguish members of that subgroup from closely related subgroups. This set is ordered from the most to the least distinguishing residues. We assume that these residues are likely responsible for functions specific to subgroup *G*. Although such a set typically includes residues with well-characterized functions, our focus is on residues of unknown functional relevance. When mapped to available structures, these distinguishing residues may readily suggest plausible hypotheses; in this respect, a BPPS analysis is informative by itself. However, SIPRIS can obtain deeper insight into and corroboration of a BPPS analysis by identifying significant overlap between BPPS-defined discriminating residues and structurally defined residue sets; we term the intersection of two such sets a **BPPS-SIPRIS cluster**. SIPRIS analysis was motivated, in part, by Karlin and Zhu's approach [113] for identifying significant clusters of residues that share physical-chemical properties.

BPPS-SIPRIS predefined clusters. The simplest BPPS-SIPRIS analysis is based on a specific, predefined structural cluster of *n* residues. This corresponds to a ball-in-urn problem, in which the BPPS-defined distinguishing residues correspond to N_1 red balls, the remaining residues to N_2 black balls, and the cluster to *n* balls drawn from the urn. The probability that at least *x* of the *n* residues are distinguishing (i.e., are "red") is given by the cumulative hypergeometric distribution:

$$P(x,n,N_1,N_2)=\left[\sum_{i=\max(x,n-N_2)}^{\min(n,N_1)}\binom{N_1}{i}\binom{N_2}{n-i}\right]\div\binom{N_1+N_2}{n}.$$

BPPS-SIPRIS optimized-clusters. Similar to BPPS-predefined clustering is choosing the optimal BPPS-structural cluster among various alternatives. To construct these, we start from a well-defined position in space, and sequentially add "structurally-adjacent" residues (variously defined, as described in Results) to generate a set of nested, structurally defined clusters. From this nested set, we select the structural cluster that optimally overlaps with the BPPS-defined residue set by applying the **Minimum Description Length (MDL)** principle [111], as described in the next section. Optimizing over different starting residues, or different numbers of discriminating residues, requires further *p*-value adjustment, for which we currently apply the overly-conservative Bonferroni correction to obtain an upper bound.

The MDL principle. To avoid overfitting BPPS-SIPRIS statistical models to observed data, we apply the **MDL** principle [111], which can be understood as formalizing Occam's Razor ("a model should not be needlessly complex"). Conceptually, this principle claims that the best among a set of alternative models is that which minimizes the description length of the model, plus the maximum-likelihood description length of the data given the model. This approach accounts for the implicit number of independent tests performed when optimizing the parameters of a model, and strikes a balance between a model's complexity and its ability to fit the data—in our case to describe biologically relevant amino acid residue patterns. More formally, a *theory* is a probability distribution over all possible sets of data, and a *model* is a parameterized set of theories. The description length of the data D given a model M, is then defined by $\mathrm{DL}(D|M) \equiv -\log \mathrm{P}(D|T)$, where T is maximum-likelihood theory contained in M (i.e. the theory which yields the greatest probability for D). The description length of the model M is defined by $\mathrm{DL}(M) \equiv \log(N)$, where N is the number of the effectively distinct theories (i.e. parameter settings) M accommodates [111]. The MDL principle aims to minimize $\mathrm{DL}(D|M) + \mathrm{DL}(M)$.

MDL applied to BPPS-SIPRIS clustering. BPPS-optimized clustering presents several mathematical challenges. Computing valid p-values requires adjusting for the multiple tests implicit in optimizing over starting residues and clusters. Also, this optimization itself may carry an implicit bias favoring small or large clusters, as outlined below.

We start with a null model in which discriminating residues (e.g., defined by BPPS) are distributed randomly throughout an entire sequence. Given a fixed number of discriminating residues, this model yields a uniform likelihood for all sets of data, and serves as a basis of comparison for likelihoods generated by an alternative model. This model divides the sequence into an initial segment of length x (which we refer to as a cluster) having m discriminating residues, and a terminal segment of length y having n discriminating residues. The model assumes discriminating residues are generated with different probabilities in the initial and terminal segments, and its maximum-likelihood theory assigns the likelihood $p=\left(m/x\right)^{m}\left(\left(x-m\right)/x\right)^{x-m}\left(n/y\right)^{n}\left(\left(y-n\right)/y\right)^{y-n}$ to the data. For a particular cut-point x, this likelihood requires choosing the discriminating-residue probabilities m/x and n/y for the initial and terminal segments, and is easily normalized for the selection of these parameters. Our aim, however, is to pick the x (i.e., cluster) that yields the greatest likelihood for the data. Applying the MDL principle requires calculating the effective number of independent tests N implicit in choosing x [110]. By treating x as a continuous as opposed to a discrete parameter, we are able to calculate its Fisher information [110], and thus N.

One subtlety is that simply choosing the cut point x yielding the greatest likelihood implicitly favors low or high values of x. This occurs because the Fisher information is greater at extreme values of x, implying that the likelihoods are more independent of one another at those values. Empirical analyses show that this bias toward large and small clusters often yields suboptimal results from a biological perspective. However, by adding an x-dependent correction, derived from the Fisher information, to our optimization, we may flatten the implicit prior associated with x [110]. Random simulation shows that analytic p-values computed using our approach fall within about 20% of empirical p-values. We still need to adjust these p-values for clusters found using different starting residues. Absent a better approach, we currently apply the simple but overly conservative Bonferroni correction [114].

Appendix 6: Statistical Tool for Analysis of Residue Couplings (STARC)

DCA and 3D contacts concurrence scores ($_{3D}S_{DC}$). The $_{3D}S_{DC}$-scores apply ICA[110] to measure, as the $-\log_{10}(p)$, the statistical significance of the correspondence between pairwise structural interactions and DC scores. Given an array of residue pairs ordered by their DC scores, we ask how well it agrees with an alternative ordering based on 3D pairwise distances. More specifically, we seek to identify an optimal initial cluster of elements of the array (defined by a cut), as measured by a relevant *p*-value. We are given an array of *L* residue pairs ordered by their DC-scores. *D* of the pairs (denoted by '1's) are separated within a reference 3D structure by ≤ *z* Å (with *z* = 3.5 Å by default) and *L* - *D* (denoted by '0's) are not.

We ask: what initial cluster, consisting of pairs up to and including a cut point *X*, contains the most surprising number *d* of '1's, and what is its probability of occurring by chance? (We term the *d* 1s in an initial cluster "left-distinguished pairs.") For *L* = 18 and *D* = 7, for example, one such array is "101101100000010001", with optimal cut point *X* = 7 (underlined), yielding *d* = 5. Since the pairs are ranked by pairwise distance, we might then represent our example array as "401603200000070005" with digits > 0 denoting the ranks of distinguished pairs. ICA ignores these ranks when choosing the optimal *X*, whereas we would prefer the *d* distinguished pairs to the left of *X* to have superior ranks (i.e., lower numbers) than those to the right.

To generalize ICA to exploit ranking information we incorporate a ball-in-urn model to calculate a ranking specific *p*-value P_b. For a specific cut point *X* that yields *d* left-distinguished pairs, we imagine first coloring red, among all *D* distinguished pairs, those *d* pairs with the smallest pairwise distances; and then recording the number *R* that are red among the left-distinguished pairs. Ideally, all the left-distinguished pairs will outrank the remaining distinguished pairs, yielding *R* = *d*, but more generally higher values of *R* are better; in the example of the previous paragraph, *D* = 7, *d* = 5 and *R* = 4. Given the null hypothesis that rankings are random, we may then use the cumulative hypergeometric distribution to calculate the probability P_b that ≥ *R* of the left-distinguished pairs are red:

$$P_b = \left[\sum_{i=R}^{d} \binom{d}{i} \binom{D-d}{d-i} \right] \div \binom{D}{d}.$$

This corresponds to drawing *d* balls from an urn containing *D* balls, of which *d* are red; note that the number of balls drawn here equals the number colored red. A low value of P_b is reported for a cut with a surprising number, among its *d* left-distinguished pairs, having the *d* smallest pairwise distances.

Before it corrects for having optimized over all possible cuts, ICA can be understood as calculating a *p*-value P_a for finding *d* distinguished pairs to the left of a cut point *X*. Because the calculation of P_a ignores ranking information, it will be independent of P_b, and these two *p*-values may therefore be combined to yield a joint *p*-value P_J [115-117] using the formula

$$P_J = P_a P_b \left(1 - \ln P_a P_b\right).$$

Low values of P_J may arise from low values of P_a, or P_b, or of both. P_J can provide a statistically stronger measure of the congruence of two orderings, here derived from DC scores and 3D distances, than does P_a alone. The *p*-values *P* we report in this paper correspond to P_J, after it has been corrected for optimization over the multiple cut points *X* considered[110].

For homomeric structures, DARC assesses the correspondence, not only between DCA scores and *internal* 3D-contacts alone (e.g., labeled as 'A' for chain A), but also between DCA scores and both *internal* and adjacent subunit *interface* 3D-contacts (e.g., labeled as ‘A:B’ for chain A and adjacent chain B). The change in $_{3D}S_{DC}$ upon inclusion of interface contacts is denoted as ΔS_{DC}. High positive values for ΔS_{DC} suggest that strong selective pressures are maintaining 3D contacts between adjacent subunits. In contrast, negative values for ΔS suggest that subunit interactions may be functionally insignificant and perhaps due to a crystallographic artifact.

BPPS and DCA or 3D contact concurrence scores ($_{DC}S_P$, $_{3D}S_P$). DARC provides a measure of statistical significance (based on ICA) for the concurrence between pairs of BPPS-defined residues and either the highest scoring DC pairs or the closest 3D-contacts. The overlap between BPPS and DCA[76] assessed in this way is often weak. Hence, DCA and BPPS are often complementary, so that combining both analyses often provides deeper insight into the relationship between protein structure and function.

Pattern residue 3D-clustering significance scores ($_{CL}S_P$). DARC applies ICA to estimate constraints tending to cluster BPPS-defined residues structurally[76]. The *L* positions of the ICA array correspond to the *L* residues within a 3D structure of the DARC query protein or of other proteins belonging to the query family. The 1s in the array correspond to a fixed number of BPPS-defined pattern residues and the 0s to the remaining residues. DARC orders array elements based on their 3D distance from a starting residue. It then determines the most significant 3D-cluster of these residues among a nested set of clusters, each centered on the starting residue. This is performed, starting with each of the BPPS-defined residues in turn, and the highest scoring 3D-cluster among these is reported along with the starting residue and the corresponding $_{CL}S_P$-score. The $_{CL}S_P$ score measures the significance of the intersection between a 3D cluster and the BPPS residue set. In addition to the strategy just described (termed “spherical expansion”), DARC allows either core expansion or hydrogen-bond-network expansion[76]. Core expansion sequentially adds the residue closest to a residue within the cluster’s “core”. This core is defined as the starting residue *R* plus all cluster residues whose distance to their k^{th} closest cluster residue is less than *R*’s distance to its k^{th} closest cluster residue (with *k*=7 by default; this was selected empirically to avoid both spherical- and tentacle-shaped clusters.) In this case, the cluster typically expands less symmetrically. Hydrogen-bond-network expansion sequentially adds a residue forming the closest sidechain-to-sidechain or sidechain-to-backbone hydrogen bond with a cluster residue.

MSA	PyMOL	
text	match	mismatch
dkyellow	yellow1	yellow0
red	red1	red0
orange	orange1	orange0
magenta	magenta1	magenta0
green	green1	green0
cyan	cyan1	cyan0
blue	blue1	blue0
purple	purple1	purple0
teal	teal1	teal0
brown	brown1	brown0
gray	gray1	gray0

Appendix 7: Contrast alignments for bacterial clamp loader subunits.

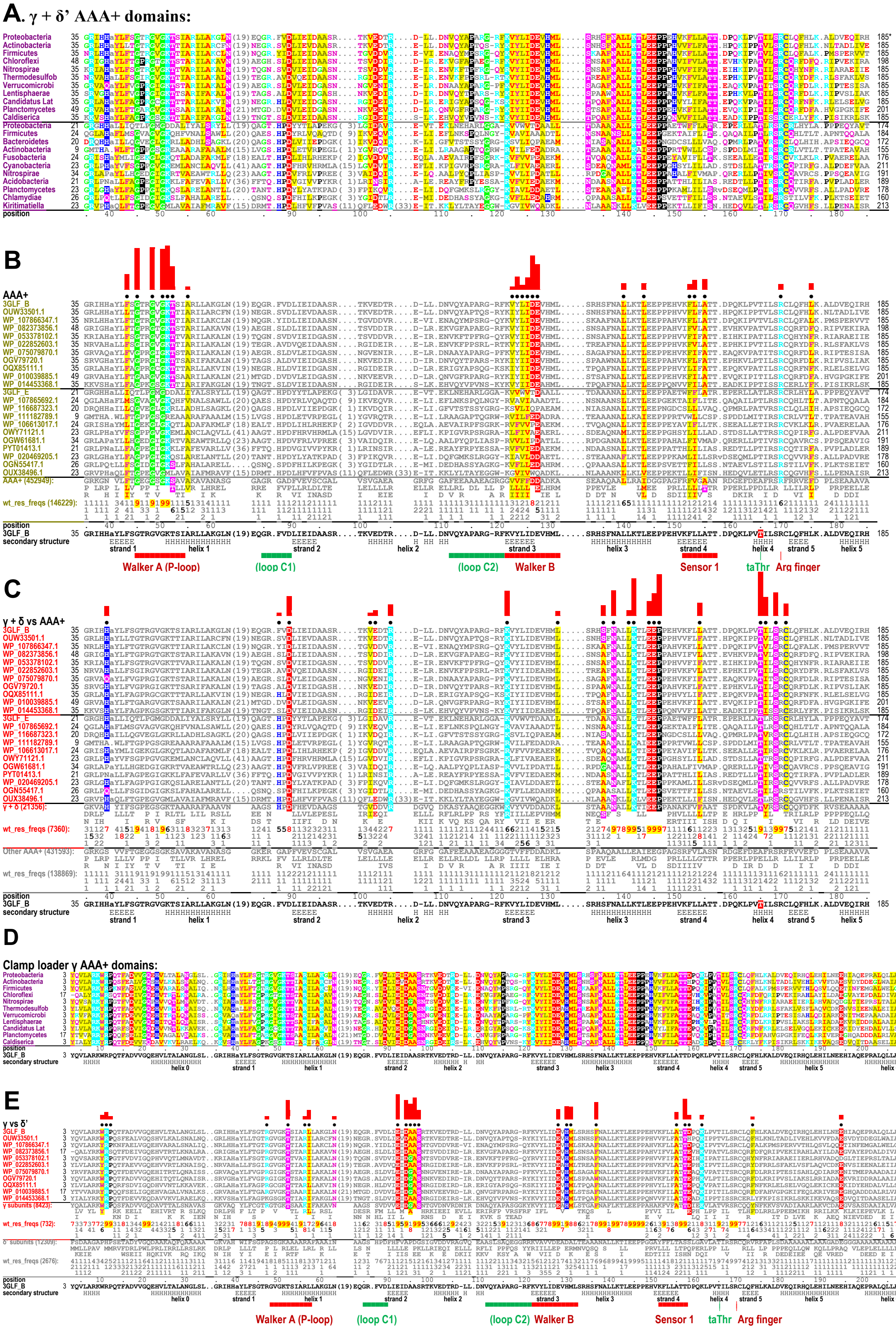


Figure 19.1. DARC-generated alignments highlighting all residues conserved in γ and δ' clamp loader proteins and residues distinctive of the AAA+ superfamily, of the γ + δ' subgroup, and of γ but not δ'. These are shown using five versions of the

same representative set of γ proteins (in panels A-E) and of δ’ proteins (in panels A-C). Residues are highlighted to indicate amino acid biochemical properties based on the following color code: red font with yellow highlight, non-polar (AVILMWFY) ; blue font with yellow highlight, cysteine (C); red, acidic (DE); cyan, basic (KR); magenta, polar (STNQ); green, glycine (G); blue, histidine (H); black, proline (P). Non-conserved positions in panels A and D and non-pattern residues in panels B,C and E are shown in gray font. The leftmost columns in panels B,C and E give the NCBI sequence identifiers; these are colored the same as the residue sidechains in **Figure 9.1**. **A**. Alignment highlighting all γ + δ’ conserved residues. Those sequences above the line within the alignment correspond to representative γ proteins from the (distinct) phyla denoted in the leftmost column; the first sequence corresponds to the E. coli γ subunit (pdb_id: 3glfB). Those sequences below the line correspond to representative δ’ proteins from distinct phyla, the first sequence of which corrsponds to the E. coli δ’ subunit (pdb_id: 3glfE), which was used as the DARC query. The positions listed at the bottom correspond to the E. coli γ subunit. **B**. BPPS contrast alignment showing the same sequences as in panel (a), but highlighting only those residues most distinctive of the AAA+ superfamily. The heights of the red bars above each highlighted column estimate the selective pressure imposed on pattern residues at that position using a semi-logarithmic scale. Directly below the aligned sequences, the characteristic AAA+ residues at each position are shown and, directly below these, corresponding frequencies are given in integer tenths; a ‘7’, for example, indicates that the corresponding residue occurs in 70-80% of the 452,949 AAA+ sequences in the alignment. Below this is shown the residue positions and sequence of the E. coli γ subunit (with the Thr 165 residue that was mutated to Val highlighted in red), and shown below these are predicted secondary structure elements (symbol: H, helix; E, strand), helix and strand designations, and AAA+ structural motifs (red font) and putative clamp binding loops C1 and C2 (green font). Secondary structure assignments were calculated for the E. coli γ subunit using DSSP [87]. **C**. BPPS contrast using the same format as in panel B to highlight those residues that most distinguish γ and δ subunits from other AAA+ proteins. **D**. DARC-generated alignment highlighting all residues conserved in γ. **E**. DARC-generated alignment highlighting residues distinguishing γ subunits from δ’ subunits. A few of these are conserved in other catalytically active AAA+ ATPases; see panel B.

1. References:

1. Buller, A.R., Brinkmann-Chen, S., Romney, D.K., Herger, M., Murciano-Calles, J. and Arnold, F.H. (2015) Directed evolution of the tryptophan synthase beta-subunit for stand-alone function recapitulates allosteric activation. *Proc Natl Acad Sci U S A*, **112**, 14599-14604.
2. Bacon, F. and Fowler, T. (1878) *Novum organum*. Clarendon Press.
3. Neuwald, A.F. and Poleksic, A. (2000) 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*, **28**, 3570-3580.
4. Neuwald, A.F. and Hirano, T. (2000) HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. *Genome Research*, **10**, 1445-1452.
5. Li, T., Chen, X., Garbutt, K.C., Zhou, P. and Zheng, N. (2006) Structure of DDB1 in complex with a paramyxovirus V protein: viral hijack of a propeller cluster in ubiquitin ligase. *Cell*, **124**, 105-117.
6. Ono, T., Losada, A., Hirano, M., Myers, M.P., Neuwald, A.F. and Hirano, T. (2003) Differential contributions of condensin I and condensin II to mitotic chromosome architecture in vertebrate cells. *Cell*, **115**, 109-121.
7. Neuwald, A.F. and Green, P. (1994) Detecting patterns in protein sequences. *J Mol Biol*, **239**, 698-712.
8. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658-1659.
9. Neuwald, A.F. and Altschul, S.F. (2016) Bayesian Top-Down Protein Sequence Alignment with Inferred Position-Specific Gap Penalties. *PLoS Comput Biol*, **12**, e1004936.
10. Neuwald, A.F. and Liu, J.S. (2004) Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. *BMC Bioinformatics*, **5**, 157.
11. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147**, 195-197.
12. Neuwald, A.F., Kolaczkowski, B.D. and Altschul, S.F. (2021) eCOMPASS: evaluative comparison of multiple protein alignments by statistical score. *Bioinformatics*.
13. Neuwald, A.F. (1997) An unexpected structural relationship between integral membrane phosphatases and soluble haloperoxidases. *Protein science : a publication of the Protein Society*, **6**, 1764-1767.
14. Neuwald, A.F. (2009) Rapid detection, classification and accurate alignment of up to a million or more related protein sequences *Bioinformatics*, **25**, 1869-1875.
15. Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res*, **43**, D222-226.
16. Neuwald, A.F., Lanczycki, C.J., Hodges, T.K. and Marchler-Bauer, A. (2020) Obtaining extremely large and accurate protein multiple sequence alignments from curated hierarchical alignments. *Database (Oxford)*, **2020**.
17. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **40**, D13-25.
18. Federhen, S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res*, **40**, D136-143.
19. Neuwald, A.F. (2011) Surveying the manifold divergence of an entire protein class for statistical clues to underlying biochemical mechanisms. *Statistical applications in genetics and molecular biology*, **10**, 36.
20. Neuwald, A.F., Kannan, N., Poleksic, A., Hata, N. and Liu, J.S. (2003) 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*, **13**, 673-692.

21. Liu, J.S. (2008) *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.

22. Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) Optimization by simulated annealing. *Science*, **220**, 671-680.

23. Neuwald, A.F. (2014) A Bayesian sampler for optimization of protein domain hierarchies. *Journal of Computational Biology*.

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

25. Kannan, N., Haste, N., Taylor, S.S. and Neuwald, A.F. (2007) The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc Natl Acad Sci U S A*, **104**, 1272-1277.

26. Kannan, N. and Neuwald, A.F. (2004) Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2{alpha}. *Protein science : a publication of the Protein Society*, **13**, 2059-2077.

27. Kannan, N. and Neuwald, A.F. (2005) Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? *J Mol Biol*, **351**, 956-972.

28. Kannan, N., Neuwald, A.F. and Taylor, S.S. (2008) Analogous regulatory sites within the alphaC-beta4 loop regions of ZAP-70 tyrosine kinase and AGC kinases. *Biochim Biophys Acta*, **1784**, 27-32.

29. Kannan, N., Wu, J., Anand, G.S., Yooseph, S., Neuwald, A.F., Venter, J.C. and Taylor, S.S. (2007) Evolution of allostery in the cyclic nucleotide binding module. *Genome Biol*, **8**, R264.

30. Neuwald, A.F. (2003) Evolutionary clues to DNA polymerase III beta clamp structural mechanisms. *Nucleic Acids Res*, **31**, 4503-4516.

31. Neuwald, A.F. (2005) Evolutionary clues to eukaryotic DNA clamp-loading mechanisms: analysis of the functional constraints imposed on replication factor C AAA+ ATPases. *Nucleic Acids Res*, **33**, 3614-3628.

32. Neuwald, A.F. (2006) Hypothesis: bacterial clamp loader ATPase activation through DNA-dependent repositioning of the catalytic base and of a trans-acting catalytic threonine. *Nucleic Acids Res*, **34**, 5280-5290.

33. Neuwald, A.F. (2007) Gα-Gβγ dissociation may be due to retraction of a buried lysine and disruption of an aromatic cluster by a GTP-sensing Arg-Trp pair. *Protein Science*, **16**, 2570-2577.

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

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

36. Hall, A. (ed.) (2000) *GTPases*. Oxford University Press.

37. Li, G. and Zhang, X.C. (2004) GTP hydrolysis mechanism of Ras-like GTPases. *J Mol Biol*, **340**, 921-932.

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

39. Zhu, G., Zhai, P., Liu, J., Terzyan, S., Li, G. and Zhang, X.C. (2004) Structural basis of Rab5-Rabaptin5 interaction in endocytosis. *Nat Struct Mol Biol*, **11**, 975-983.

40. Vetter, I.R. and Wittinghofer, A. (2001) The guanine nucleotide-binding switch in three dimensions. *Science*, **294**, 1299-1304.

41. Thomas, C., Fricke, I., Scrima, A., Berken, A. and Wittinghofer, A. (2007) Structural evidence for a common intermediate in small G protein-GEF reactions. *Mol Cell*, **25**, 141-149.

42. Gasper, R., Thomas, C., Ahmadian, M.R. and Wittinghofer, A. (2008) The role of the conserved switch II glutamate in guanine nucleotide exchange factor-mediated nucleotide exchange of GTP-binding proteins. *J Mol Biol*, **379**, 51-63.

43. Chavas, L.M., Torii, S., Kamikubo, H., Kawasaki, M., Ihara, K., Kato, R., Kataoka, M., Izumi, T. and Wakatsuki, S. (2007) Structure of the small GTPase Rab27b shows an unexpected swapped dimer. *Acta Crystallogr D Biol Crystallogr*, **63**, 769-779.
44. Merkel, J.S. and Regan, L. (1998) Aromatic rescue of glycine in beta sheets. *Fold Des*, **3**, 449-455.
45. Renault, L., Kuhlmann, J., Henkel, A. and Wittinghofer, A. (2001) Structural basis for guanine nucleotide exchange on Ran by the regulator of chromosome condensation (RCC1). *Cell*, **105**, 245-255.
46. Neuwald, A.F. (2006) Bayesian shadows of molecular mechanisms cast in the light of evolution. *Trends Biochem Sciences*, **31**, 374-382.
47. Naiki, T., Kondo, T., Nakada, D., Matsumoto, K. and Sugimoto, K. (2001) Chl12 (Ctf18) forms a novel replication factor C-related complex and functions redundantly with Rad24 in the DNA replication checkpoint pathway. *Mol Cell Biol*, **21**, 5838-5845.
48. Neuwald, A.F. (2016) Gleaning structural and functional information from correlations in protein multiple sequence alignments. *Curr Opin Struct Biol*, **38**, 1-8.
49. Neuwald, A.F. and Altschul, S.F. (2016) Inference of Functionally-Relevant N-acetyltransferase Residues Based on Statistical Correlations. *PLoS Comput Biol*, **12**, e1005294.
50. Dorfmueller, H.C., Fang, W., Rao, F.V., Blair, D.E., Attrill, H. and van Aalten, D.M. (2012) Structural and biochemical characterization of a trapped coenzyme A adduct of Caenorhabditis elegans glucosamine-6-phosphate N-acetyltransferase 1. *Acta Crystallogr D Biol Crystallogr*, **68**, 1019-1029.
51. Bosch, D.E., Wittchen, E.S., Qiu, C., Burridge, K. and Siderovski, D.P. (2011) Unique structural and nucleotide exchange features of the Rho1 GTPase of Entamoeba histolytica. *J Biol Chem*, **286**, 39236-39246.
52. Eathiraj, S., Pan, X., Ritacco, C. and Lambright, D.G. (2005) Structural basis of family-wide Rab GTPase recognition by rabenosyn-5. *Nature*, **436**, 415-419.
53. Neuwald, A.F., Aravind, L. and Altschul, S.F. (2018) Inferring joint sequence-structural determinants of protein functional specificity. *Elife*, **7**.
54. Freudenthal, B.D., Beard, W.A., Cuneo, M.J., Dyrkheeva, N.S. and Wilson, S.H. (2015) Capturing snapshots of APE1 processing DNA damage. *Nat Struct Mol Biol*, **22**, 924-931.
55. Mol, C.D., Izumi, T., Mitra, S. and Tainer, J.A. (2000) DNA-bound structures and mutants reveal abasic DNA binding by APE1 and DNA repair coordination [corrected]. *Nature*, **403**, 451-456.
56. Qu, J., Liu, G.H., Huang, B. and Chen, C. (2007) Nitric oxide controls nuclear export of APE1/Ref-1 through S-nitrosation of cysteines 93 and 310. *Nucleic Acids Res*, **35**, 2522-2532.
57. Tresaugues, L., Silvander, C., Flodin, S., Welin, M., Nyman, T., Graslund, S., Hammarstrom, M., Berglund, H. and Nordlund, P. (2014) Structural basis for phosphoinositide substrate recognition, catalysis, and membrane interactions in human inositol polyphosphate 5-phosphatases. *Structure*, **22**, 744-755.
58. Speed, C.J., Matzaris, M., Bird, P.I. and Mitchell, C.A. (1995) Tissue distribution and intracellular localisation of the 75-kDa inositol polyphosphate 5-phosphatase. *Eur J Biochem*, **234**, 216-224.
59. Mills, S.J., Silvander, C., Cozier, G., Tresaugues, L., Nordlund, P. and Potter, B.V. (2016) Crystal Structures of Type-II Inositol Polyphosphate 5-Phosphatase INPP5B with Synthetic Inositol Polyphosphate Surrogates Reveal New Mechanistic Insights for the Inositol 5-Phosphatase Family. *Biochemistry*, **55**, 1384-1397.
60. de Juan, D., Pazos, F. and Valencia, A. (2013) Emerging methods in protein co-evolution. *Nat Rev Genet*, **14**, 249-261.
61. Cocco, S., Monasson, R. and Weigt, M. (2013) From principal component to direct coupling analysis of coevolution in proteins: low-eigenvalue modes are needed for structure prediction. *PLoS Comput Biol*, **9**, e1003176.

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

63. Hopf, T.A., Colwell, L.J., Sheridan, R., Rost, B., Sander, C. and Marks, D.S. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607-1621.

64. Morcos, F., Hwa, T., Onuchic, J.N. and Weigt, M. (2014) Direct coupling analysis for protein contact prediction. *Methods Mol Biol*, **1137**, 55-70.

65. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T. and Weigt, M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A*, **108**, E1293-1301.

66. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. and Aurell, E. (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys*, **87**, 012707.

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

68. Stein, R.R., Marks, D.S. and Sander, C. (2015) Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Comput Biol*, **11**, e1004182.

69. Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R. and Sander, C. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.

70. Hopf, T.A., Morinaga, S., Ihara, S., Touhara, K., Marks, D.S. and Benton, R. (2015) Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat Commun*, **6**, 6077.

71. Dwyer, R.S., Ricci, D.P., Colwell, L.J., Silhavy, T.J. and Wingreen, N.S. (2013) Predicting functionally informative mutations in Escherichia coli BamA using evolutionary covariance analysis. *Genetics*, **195**, 443-455.

72. Espada, R., Parra, R.G., Mora, T., Walczak, A.M. and Ferreiro, D.U. (2015) Capturing coevolutionary signals inrepeat proteins. *BMC Bioinformatics*, **16**, 207.

73. Hopf, T.A., Scharfe, C.P., Rodrigues, J.P., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M. and Marks, D.S. (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, **3**.

74. dos Santos, R.N., Morcos, F., Jana, B., Andricopulo, A.D. and Onuchic, J.N. (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. *Scientific reports*, **5**, 13652.

75. Cheng, R.R., Morcos, F., Levine, H. and Onuchic, J.N. (2014) Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc Natl Acad Sci U S A*, **111**, E563-571.

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

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

78. Jones, D.T., Buchan, D.W., Cozzetto, D. and Pontil, M. (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184-190.

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

80. Tondnevis, F., Dudenhausen, E.E., Miller, A.M., McKenna, R., Altschul, S.F., Bloom, L.B. and Neuwald, A.F. (2020) Deep Analysis of Residue Constraints (DARC): identifying determinants of protein functional specificity. *Scientific reports*, **10**, 1691.

81. DeLano, W.L. (2002). DeLano Scientific, Palo Alto, CA, USA.

82. Schrodinger, LLC. (2010).

83. Hedglin, M., Kumar, R. and Benkovic, S.J. (2013) Replication clamps and clamp loaders. *Cold Spring Harb Perspect Biol*, **5**, a010165.

84. Kelch, B.A., Makino, D.L., O'Donnell, M. and Kuriyan, J. (2012) Clamp loader ATPases and the evolution of DNA replication machinery. *BMC Biol*, **10**, 34.

85. Indiani, C. and O'Donnell, M. (2006) The replication clamp-loading machine at work in the three domains of life. *Nat Rev Mol Cell Biol*, **7**, 751-761.

86. Simonetta, K.R., Kazmirski, S.L., Goedken, E.R., Cantor, A.J., Kelch, B.A., McNally, R., Seyedin, S.N., Makino, D.L., O'Donnell, M. and Kuriyan, J. (2009) The mechanism of ATP-dependent primer-template recognition by a clamp loader complex. *Cell*, **137**, 659-671.

87. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

88. Neuwald, A.F., Yang, H. and Tracy Nixon, B. (2022) SPARC: Structural properties associated with residue constraints. *Comput Struct Biotechnol J*, **20**, 1702-1715.

89. Shen, C., Lu, A., Xie, W.J., Ruan, J., Negro, R., Egelman, E.H., Fu, T.M. and Wu, H. (2019) Molecular mechanism for NLRP6 inflammasome assembly and activation. *Proc Natl Acad Sci U S A*, **116**, 2052-2057.

90. Bae, J.Y. and Park, H.H. (2011) Crystal structure of NALP3 protein pyrin domain (PYD) and its implications in inflammasome assembly. *J Biol Chem*, **286**, 39528-39536.

91. Wang, Y., Lamim Ribeiro, J.M. and Tiwary, P. (2020) Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current Opinion in Structural Biology*, **61**, 139-145.

92. Hildebrand, P.W., Rose, A.S. and Tiemann, J.K.S. (2019) Bringing Molecular Dynamics Simulation Data into View. *Trends in Biochemical Sciences*, **44**, 902-913.

93. Bowerman, S. and Wereszczynski, J. (2016) In Voth, G. A. (ed.), *Methods in Enzymology*. Academic Press, Vol. 578, pp. 429-447.

94. Noé, F. and Clementi, C. (2017) Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Current Opinion in Structural Biology*, **43**, 141-147.

95. Jin, Y., Johannissen, L.O. and Hay, S. (2021) Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins: Structure, Function, and Bioinformatics*, **89**, 915-921.

96. Michaud-Agrawal, N., Denning, E.J., Woolf, T.B. and Beckstein, O. (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem*, **32**, 2319-2327.

97. Carrillo-Tripp, M., Alvarez-Rivera, L., Lara-Ramírez, O.I., Becerra-Toledo, F.J., Vega-Ramírez, A., Quijas-Valades, E., González-Zavala, E., González-Vázquez, J.C., García-Vieyra, J., Santoyo-Rivera, N.B. *et al.* (2018) HTMoL: full-stack solution for remote access, visualization, and analysis of molecular dynamics trajectory data. *J Comput Aided Mol Des*, **32**, 869-876.

98. Bush, M. and Dixon, R. (2012) The role of bacterial enhancer binding proteins as specialized activators of sigma54-dependent transcription. *Microbiol Mol Biol Rev*, **76**, 497-529.

99. Gao, F., Danson, A.E., Ye, F., Jovanovic, M., Buck, M. and Zhang, X. (2020) Bacterial Enhancer Binding Proteins-AAA(+) Proteins in Transcription Activation. *Biomolecules*, **10**.

100. Sysoeva, T.A., Chowdhury, S., Guo, L. and Nixon, B.T. (2013) Nucleotide-induced asymmetry within ATPase activator ring drives σ54-RNAP interaction and ATP hydrolysis. *Genes Dev*, **27**, 2500-2511.

101. Chen, B., Sysoeva, T.A., Chowdhury, S., Guo, L., De Carlo, S., Hanson, J.A., Yang, H. and Nixon, B.T. (2010) Engagement of arginine finger to ATP triggers large conformational changes in NtrC1 AAA+ ATPase for remodeling bacterial RNA polymerase. *Structure*, **18**, 1420-1430.

102. Brown, M., Hughey, R., Krogh, A., Mian, I.S., Sjolander, K. and Haussler, D. (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Ismb*, **1**, 47-55.

103. Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S. and Haussler, D. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci*, **12**, 327-345.

104. Nguyen, V.A., Boyd-Graber, J. and Altschul, S.F. (2013) Dirichlet mixtures, the Dirichlet process, and the structure of protein space. *Journal of computational biology : a journal of computational molecular cell biology*, **20**, 1-18.

105. Henikoff, S. and Henikoff, J.G. (1994) Position-based sequence weights. *J Mol Biol*, **243**, 574-578.

106. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

107. Knuth, D.E. (1997) *The Art of Computer Programming*. 3rd ed.

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

109. Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D. *et al.* (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745-749.

110. Altschul, S.F. and Neuwald, A.F. (2018) Initial Cluster Analysis. *Journal of computational biology : a journal of computational molecular cell biology*, **25**, 121-129.

111. Grunwald, P.D. (2007) *The minimum description length principle*. MIT Press, Boston.

112. Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

113. Karlin, S. and Zhu, Z.Y. (1996) Characterizations of diverse residue clusters in protein three-dimensional structures. *Proc Natl Acad Sci U S A*, **93**, 8344-8349.

114. Bonferroni, C.E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3-62.

115. Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48-54.

116. Fisher, R.A. (1954) *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, Scotland.

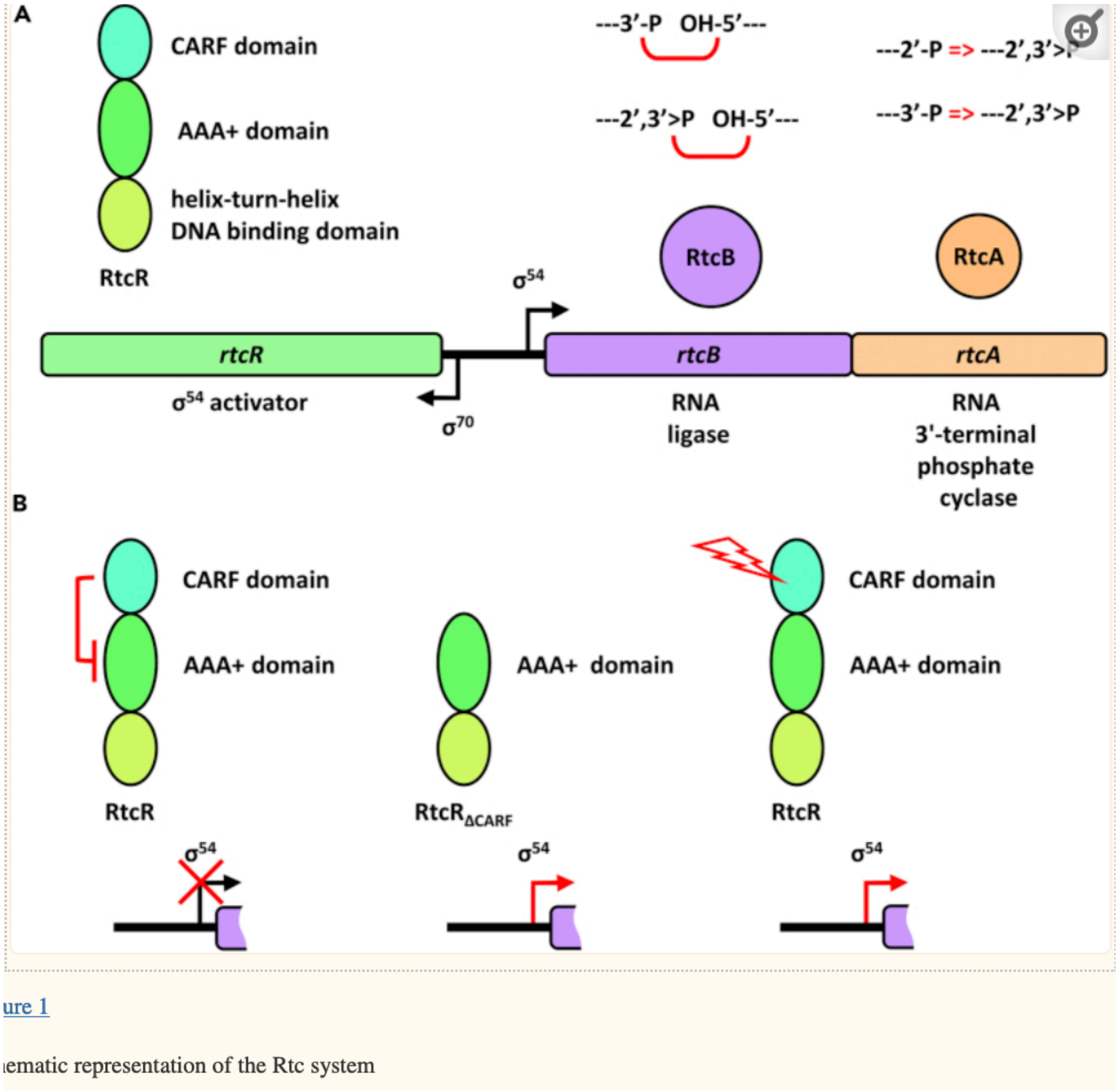
117. Yu, Y.K., Gertz, E.M., Agarwala, R., Schaffer, A.A. and Altschul, S.F. (2006) Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. *Nucleic Acids Res*, **34**, 5966-5973.

2. Acknowledgement:

This document describes a suite of programs that were developed in collaboration with statisticians Stephen F. Altschul (NCBI, NIH)(Gibbs sampler, GISMO, BPPS, STARC, SIPRIS, DARC, cCOMPASS), Jun S. Liu (Harvard)(Gibbs sampler, GISMO, BPPS), and Charles E. Lawrence (Brown University)(Gibbs sampler).

NOTES:

RtcR bEBP: CARF relaces the receiver domain.



ure 1

ematic representation of the Rtc system

For example,

- **Gibbs Sampler**
- **PSIRD**
- **GISMO**
- **BPPS**
    - edit_hpt
    - **Convert_hpt = mgs2sma?**
    - ~~**bpps2vsi = sarp – DONE.**~~
    - ~~**vsi2pml (bpps P …) DONE.**~~
- **SPARC**
- **STARC**
    - **Dca**
    - **cma2aln**
- **SIPRIS**
- **LAPIS**
- **eCOMPASS**
- **DARC**
- **MAPGAPS**
    - ~~**mgs2sma**~~
- **twkcma**
    - Unrelatedseq
    - Gor_driver (2ndary struct prediction)
    - Goscan?
    - purgecma
    - editcma
    - convert (ConvertMSA)
        - Add sto2fa
        - selex2cma
        - msa2clustalw
        - cma2sto (see twkcma X)
- Twkpdb
    - Superimpose
    - Vsi2pml
    - Pdb4vsi
    - FRP2vsi
    - get_pdb
- Twkseq
    - Purge
    - repset
    - Cdhit
    - Rpts
    - Swaln, swalign
    - Weights
    - Asset ???
    - Rmseq
    - split
    - AddPhyla

        - tax2msa (old code)
- gblastpgp
- <u>others:</u>
- GOSCAN
- Protein
- probe