The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008
Copyright © 2008 ORSC & APORC, pp. 395–403

Structure-based functional annotation of protein sequences guided by comparative models

Daron. M. Standley[1,*] Akira R. Kinjo[2] Miesko Lis[3]
Mark van der Giezen[4] Haruki Nakamura[2]

[1]Systems Immunology Laboratory, Immunology Frontier Research Center,
Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan.
[2]Research Center of Structural and Functional Proteomics, Institute for Protein Research,
Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan
[3]MIT Computer Science and Artificial Intelligence Laboratory,
32 Vassar Street, Cambridge, MA 02139, USA
[4]Centre for Eukaryotic Evolutionary Microbiology, School of Biosciences, University of Exeter,
Stocker Road, Exeter EX4 4QD, UK.

Abstract A strategy for functionally annotating protein sequences using sequence and predicted structural information is proposed. First, structural models are built using standard, web-based tools. The models are then annotated using the sequence and structure-based annotation method SeSAW, and the structure based ligand binding site alignment method GIRAF. Annotations for several sequences (ATPase-like domains 1 and 2 from mouse RIG-I, the TIR domain from mouse Toll-like receptor 9, and alternative oxidase from *Blastocystis*) are made and compared with known functional information. We find that the annotations in general make functional sense and provide more specific information than would be available from a purely sequence-based approach.

1 Introduction

High-throughput gene sequencing projects have revealed the complete genomes of over 180 organisms and are currently engaged in sequencing numerous others. To make biological sense of such large volumes of data, it is necessary to compare the protein sequences with those of proteins that have been biochemically characterized. Structural genomics (SG) efforts facilitate such comparisons by determining the structures for a large number of protein sequences, but most SG targets have not been functionally characterized. We have previously shown that functional details can nevertheless be inferred by sequence and structural comparison to other structures for which the functions are known [1]. This suggests that in order to learn the function of a protein sequence it may be useful to build a 3D model using a remote homolog as a structural template, even if the template has no functional annotation. The model can often then be partially annotated using a

[*] Contact: standley@IFReC.osaka-u.ac.jp

combination of sequence and structural comparison to known folds. In the second step it is crucial to optimize the integration of sequence and structural information in order to identify the most relevant functional template as well as the most important residue positions within this template.

Recently, we developed two structural alignment methods that facilitate such sequence or structure based functional annotations. Sequence-derived Structural Alignment Weights (SeSAW) is a method that optimally integrates sequence and structural information at the level of individual residues in order to identify conserved motifs [1]. Geometric Indexing with Refined Alignment Finder (GIRAF) is a template-based method that matches known ligand binding sites with similar sites in the query at the atomic level [2]. Both methods were originally developed for annotation of experimentally determined structures. Here we extend these methods to the functional annotation of sequences by using a structural model as an intermediate step. We demonstrate the utility of this approach using several examples where experimental structures for the sequences in question are not yet available, but some functional annotations exist.

2 Methods

2.1 Structural modeling

The full-length query sequence was first submitted to the HHpred server [3] in order to obtain approximate domain boundaries. The sequence segments corresponding to these domains were subsequently submitted to the HHpred server again using default settings. MODELLER [4] was used to build single-template models using the highest scoring template in each case.

2.2 GIRAF Queries

GIRAF is a method for finding ligand binding sites that are structurally similar at the atomic level to sub-structures in a query protein [2]. To define a local coordinate system, we use the Delaunay tessellation of protein atoms. Each Delaunay tetrahedron is characterized by its edge lengths, volume, and the compositions of atom types in the direction of each of the four faces. Tetrahedra thus obtained are used to describe the atomic coordinates of ligand binding sites, which are saved and indexed in a relational database. The use of the relational database offers an advantage for rapidly handling vast amounts of data for ligand binding sites. There are currently more than 180,000 such sites in the PDB.

A search is carried out in two phases. First, the query structures are transformed to local coordinate systems in the same manner as the templates. Owing to the database system and indexing, potential matches can be found with essentially one SQL expression (although some modifications are necessary for performance optimization). Second, potential candidates thus obtained are then subject to alignment refinement, which is carried out as follows: (1) The template and the query are superimposed based on the local coordinates defined by their respective tetrahedra; (2) Potential atom-atom correspondences (possibly many-to-many) are identified with a given distance cutoff, which yields a bipartite graph; (3) The Hungarian algorithm is applied to the bipartite graph to obtain the best alignment; (4) Based on the alignment, a new optimal superposition is calculated; (5)

The steps (2) to (4) are iterated until convergence. A statistical model for estimating the significance of a match is also introduced.

2.3 SeSAW Queries

SeSAW [5] is a functional annotation server that identifies conserved sequence and structural motifs in a query protein. SeSAW first identifies all entries in the Protein Data Bank (PDB) that are structurally related to the query. The functional significance of each structural match (template) is then assessed by profile-profile sequence comparisons anchored by the structure-based sequence alignments. Functional sites, when available, are then mapped from the templates onto the query-template alignments. A list of the templates, sorted by their functional significance is returned, with links to both the annotated alignments and the 3D structural superpositions.

The SeSAW server maintains a database of all-against-all structural alignments for a sequence-representative set of structural domains. This database is kept current with weekly updates. Functional information is extracted from UniProt and literature sources on a weekly basis. This information is then mapped onto the residue positions of each PDB entry. Structure-based sequence alignments indicating the per-residue SeSAW score are displayed using the Jalview java-based alignment applet [6]. Functional annotations, when available, and secondary structure assignments are indicated as well. Structural superpositions are visualized using the *jV* molecular graphics viewer [7]. A panel lists the aligned residue pairs, ranked by the per-residue SeSAW score. Each residue pair can be seen in stick form in the *jV* window or hidden. This feature is important because many high-scoring residue pairs are structurally important (e.g. proline, glycine, cystine, and other hydrophobic residues) and should be hidden in order to more easily locate putative functional sites (which tend to include more chemically active residues).

3 Results and discussion

Results for SeSAW queries are shown in figures 1-5. Results for GIRAF queries are summarized in Table 1. Individual results are discussed below.

3.1 Helicase domain from mouse RIG-I

The sequence for the helicase domain from mouse RIG-I was submitted to the HHpred profile-profile alignment server. The top structural match was to the Hef helicase domain from *Pyrococcus furiosus* (PDB identifier 1wp9). Examination of the alignment and template structure revealed that the fold consisted of two ATPase-like domains connected by a mostly helical domain. Although the predicted secondary structure for the RIG-I sequence segment that aligned to the helical domain was also predicted to be helical, the sequences did not match well in this region. For this reason, the RIG-I helicase domain sequence was split into three segments corresponding to the two ATPase-like domains and the helical region. The arrangement of domains is illustrated in figure 1. The three sub-domain sequences were resubmitted to the HHpred server. Subsequently, a 44% identical sequence homolog, human MDA5 helicase domain (PDB identifier 3b6e), was found for the first ATPase domain. A more remote 19% identical homolog, the ATP-dependent RNA helicase from fruit fly (PDB identifier 2db3) was located for the second ATPase domain. Both of these templates were SG targets and a primary literature reference existed only for 2db3. The first ATPase-like domain contains a DExH ATP-binding

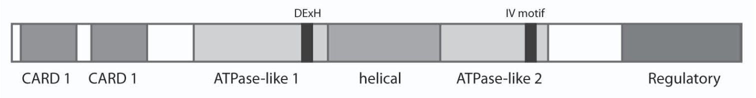


Figure 1: **Domain structure of RIG-I.** The figure shows the locations of the two ATPase-like domains in the context of the while RIG-I protein sequence. The two ATPase-like domains, along with the helical domain make up the helicase functional domain.

motif. Thus, it was not surprising to see that these resides scored highly in the SeSAW alignment (figure 2B). More interestingly, a conserved lysine residue (here denoted K30) that also scored highly, was found to form a salt bridge to the aspartic acid (denoted as D132) in the DExH motif (figure 2C). Yoneyama and Fujita noted that mutating this lysine to alanine, rendered RIG-I a dominant inhibitor [8] The GIRAF query returned a sodium binding site located in a loop formed near K30 and another residue that scored highly by SeSAW (P25).

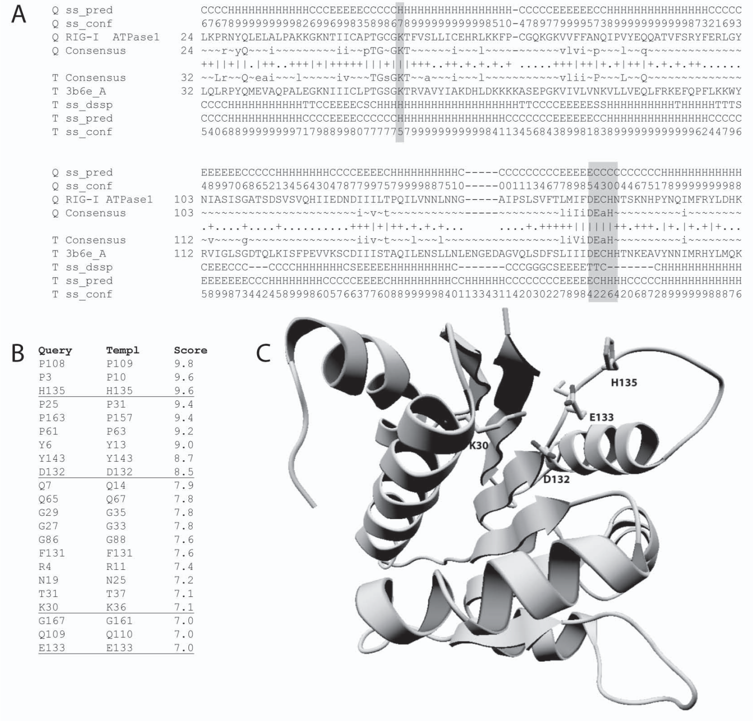
The second ATPase-like domain does not contain a DExH ATP binding motif. The highest-scoring residues in the SeSAW alignment are actually on the opposite side of the protein, and include a conserved arginine (R131) that forms a salt-bridge with a conserved aspartic acid (D104), as shown in figure 3C. The surface patch where the arginine is located maps to the helicase motif IV in the Hef helicase from *Pyrococcus furiosus* [9]. The GIRAF query indicated a sulfate binding site at the N-terminal helix, which is where the two ATPase domains are joined by the structurally uncharacterized alpha-domain. The significance of this binding site is not known. These observations suggest that the first domain might act as an ATPase, while the second domain probably has another function, such as nucleotide binding. RIG-I is known to sense double-stranded viral RNA as part of the innate immune response [8].

3.2 Toll-like receptor-9 TIR domain from mouse

The model of the Toll/interleukin-1 receptor from mouse Toll-like receptor 9 (TLR-9 TIR domain) was built on human TLR-1 (1fyv) with a sequence identity of 20%. The highest-scoring residue pairs in the SeSAW alignment included a number of conserved hydrophobic residues (W16, W125, Y110, F135, W136), along with prolines (P45,P103,P126) that are characteristic of TIR domains. Moreover, five residues in a conserved patch that includes the ‘cBB loop,’ thought to be involved in receptor signaling [10], score highly as well, as indicated in figure 4B. The GIRAF query returned no hits, which is not unexpected because the TIR domain is involved in protein-protein interactions rather than ligand interactions. Differences between the TIR domain in TLR-9 and TLR-1 are consistent with a different signal cascades initiated by these two innate immune response receptors: TLR-9 detects unmethylated CpG DNA and TLR-1 detects lipoproteins derived from bacteria.

3.3 Alternative oxidase from *Blastocystis*

A model of alternative oxidase (AOX) from the intestinal parasite *Blastocystis* was built using bacterioferritin from *Mycobacterium smegmatis* (3bkn) as a template. The sequence identity was only 16% but nevertheless statistically significant due to the matching

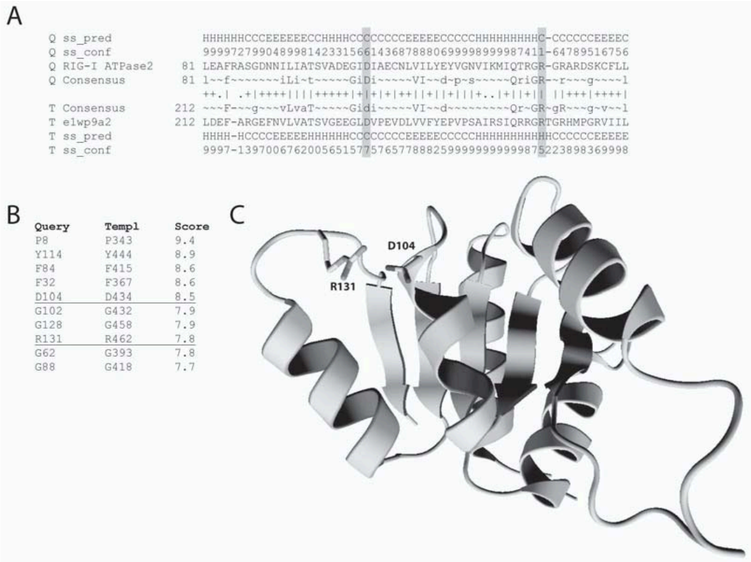


```
Q ss_pred                 CCCCHHHHHHHHHHHHHCCCEEEEECCCCCHHHHHHHHHHHHHHHHH-CCCCCEEEEEECCHHHHHHHHHHHHHHHHCCCCC
Q ss_conf                 676789999999998269969983589867899999999999998510-47897799995738999999999999987321693
Q RIG-I  ATPase1   24 LKPRNYQLELALPAKKGKNTIICAPTGCGKTFVSLLICEHRLKKFP-CGQKGKVVFFANQIPVYEQQATVFSRYFERLGY
Q Consensus       24 ~~~r~yQ~~~i~~~~~~~~~~ii~~pTG~GKT~~~~~~i~~~l~~~~-~~~~~~vlvi~p~~~l~~q~~~~~~~~~~~~~~
                          ++||+||.|++..+++|+|+++++|||+|||+++..++.+.+.+.+ .....+++|++|+.+|++|+.+.+.++......
T Consensus       32 ~~Lr~~Q~eai~~~l~~~~~iv~~~TGsGKT~~a~~~i~~~l~~~~~~~~~~~~~~vlii~P~~~L~~Q~~~~~~~~~~~~~~
T 3b6e_A         32 LQLRPYQMEVAQPALEGKNIIICLPTGSGKTRVAVYIAKDHLDKKKKASEPGKVIVLVNKVLLVEQLFRKEFQPFLKKWY
T ss_dssp                 CCCCHHHHHHHHHHHTTCCEEEECSCHHHHHHHHHHHHHHHHHHHHHHTTCCCCEEEEESSHHHHHHHHHHTHHHHHTTTS
T ss_pred                 CCCCHHHHHHHHHHHCCCCEEEECCCCCHHHHHHHHHHHHHHHHHCCCCCCCEEEEEECCHHHHHHHHHHHHHHHHCCCCC
T ss_conf                 540688999999997179889980777757999999999999841134568438998183899999999999996244796

Q ss_pred                 EEEEEECCCCCHHHHHHHHCCCCEEEECHHHHHHHHHHC-----CCCCCCCCCEEEEECCCCCCCCCHHHHHHHHHHHH
Q ss_conf                 489970686521345643047877997579999887510-----00111346778985430044675178999999988
Q RIG-I ATPase1  103 NIASISGATSDSVSVQHIIEDNDIIILTPQILVNNLNNG-----AIPSLSVFTLMIFDECHNTSKNHPYNQIMFRYLDHK
Q Consensus      103 ~~~~~~~~~~~~~~~~~~~~~~~~i~v~t~~~~~~~~~~~-----~~~~~~~~~~liIiDEaH~~~~~~~~~~~i~~~~~~~~
                          .+....+....+..........+++|+|++.+.+.+...     ....+.+++++|||||++.+++.|+.|+.+|+...
T Consensus      112 ~v~~~~g~~~~~~~~~~~~~~~~~~iiv~t~~~l~~~l~~~~~~~~~~~~~~~~~liViDEaH~~~~~~~~~~~i~~~~~~~~
T 3b6e_A        112 RVIGLSGDTQLKISFPEVVKSCDIIISTAQILENSLLNLENGEDAGVQLSDFSLIIIDECHHTNKEAVYNNIMRHYLMQK
T ss_dssp                 CEEECCC---CCCCHHHHHHHCSEEEEEHHHHHHHHHC-------CCCGGGCSEEEETTC-------CHHHHHHHHHHHH
T ss_pred                 EEEEEEECCCHHHHHHHHCCCCCEEEECHHHHHHHHHCCHHHHHHCCCCCCCCEEEEECHHHCCCCCHHHHHHHHHHHH
T ss_conf                 589987344245899986057663776088999999840113343114203022789842264206872899999988876
```

Query	Templ	Score
P108	P109	9.8
P3	P10	9.6
H135	H135	9.6
P25	P31	9.4
P163	P157	9.4
P61	P63	9.2
Y6	Y13	9.0
Y143	Y143	8.7
D132	D132	8.5
Q7	Q14	7.9
Q65	Q67	7.8
G29	G35	7.8
G27	G33	7.8
G86	G88	7.6
F131	F131	7.6
R4	R11	7.4
N19	N25	7.2
T31	T37	7.1
K30	K36	7.1
G167	G161	7.0
Q109	Q110	7.0
E133	E133	7.0

Figure 2: **Annotation of RIG-I ATPase-like domain 1.** The figure shows A) a fragment of the raw HHpred alignment to the top conserved domain hit; B) the top-scoring residue pairs between the query and 3b6e according to SeSAW, with key residues underlined; C) the SeSAW structural model with the key residues shown as sticks.

of iron coordinating glutamic acid and histidine residues and secondary structure. The SeSAW query clearly identified these important residues, and also ranked a number of aspartic acids very highly. The significance of these aspartic acid residues is not known, but the iron-ligating residues that consist of two E...HxxE motifs, as well as the existence of a key tyrosine and trypophan residue, have been established [11]. The two iron-binding motifs, along with the conserved tyrosine align and rank highly in the SeSAw alignment. As figure 5C shows, these residues form a tight cluster. The spatial clustering gives us more confidence in our model and annotation than a purely sequence-based alignment and annotation would. The GIRAF query returned a very strong match to a designed protein that binds dimethyl sulfide at the same location as the expected Fe^{2+} binding site (PDB identifier 1jm0). The next-best match was to the Fe^{2+} binding site in bacterioferritin. These two hits were much more significant than any of the other GIRAF queries, consistent with a precise Fe^{2+} binding signature, showing the utility of the atom-based GIRAF alignment in ligand binding site identification.



```
Q ss_pred              HHHHHHCCCEEEEEEECCHHHHCCCCCCCCEEEEECCCCHHHHHHHHHC-CCCCCCEEEEC
Q ss_conf              99997279904899814233156614368788806999989999987411-64789516756
Q RIG-I ATPase2     81 LEAFRASGDNNILIATSVADEGIDIAECNLVILYEYVGNVIKMIQTRGR-GRARDSKCFLL
Q Consensus         81 l~~f~~~~~~~iLi~t~~~~~GiDi~~~~~VI~~d~p~s~~~~~QriGR~~r~~~g~~~~l
                       ++.| .+|++++||||+++++|+|+|+|++||+||+|||+..|+||+|| ||+++|++++|

T Consensus        212 ~~~F~~~g~~~vLvaT~~~~~Gidi~~~~~VI~~d~~~~~~~~~~Qr~GR~gR~~~g~v~~l
T e1wp9a2          212 LDEF-ARGEFNVLVATSVGEEGLDVPEVDLVVFYEPVPSAIRSIQRRGRTGRHMPGRVIIL
T ss_pred                HHHH-HCCCCEEEEEHHHHHHCCCCCCCCEEEEECCCCHHHHHHHHHHHCCCCCCEEEEE
T ss_conf                9997-139700676200565157757657788825999999999999875223898369998
```

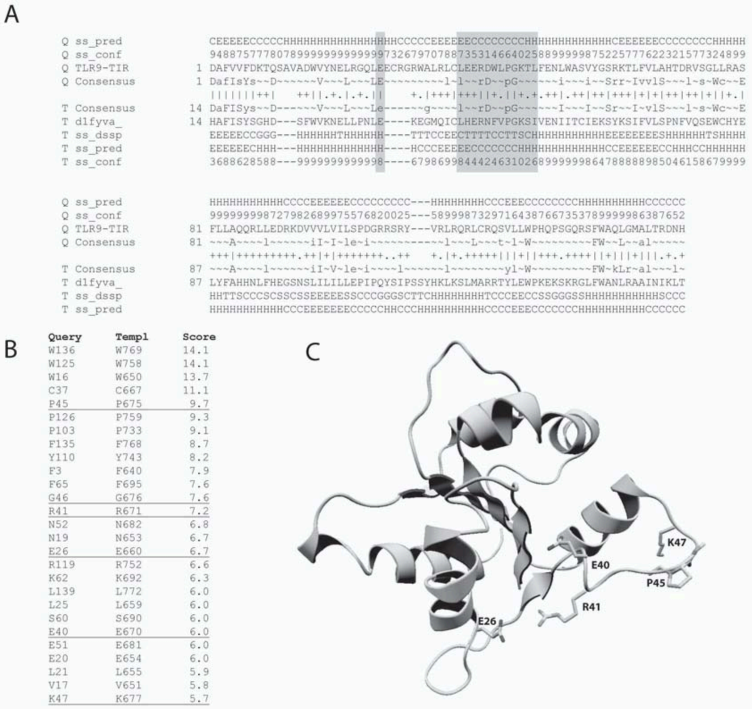
Query	Templ	Score
P8	P343	9.4
Y114	Y444	8.9
F84	F415	8.6
F32	F367	8.6
D104	D434	8.5
G102	G432	7.9
G128	G458	7.9
R131	R462	7.8
G62	G393	7.8
G88	G418	7.7

Figure 3: **Annotation of RIG-I ATPase-like domain 2.** The figure shows A) the raw HHpred alignment to the top conserved domain hit; B) the top-scoring residue pairs between the query and 2db3 according to SeSAW, with key residues underlined; C) the SeSAW structural model with the key residues shown as sticks.

4 Conclusions

We have demonstrated here that structural modeling can be used as an intermediate step to map functional sites onto sequences without an experimentally-determined structure. The functional annotation methods used were developed by the authors, but alternate structure-based functional annotation tools [12, 13] could just as easily have been used. The two criteria required for this approach to work are 1) an accurate structural model must be built for at least the regions of interest and 2) distant homologs to the template that have been functionally characterized must exist. These criteria are less strict than those required for a direct sequence-based annotation approach. To demonstrate this, we have included the raw HHpred alignments in figures 2A-5A. As can be seen, many exact or high-scoring matches exist in each alignment, so identifying the functionally important residues from this set would be difficult without putting the matches into a structural context. SeSAW and GIRAF automate the process of identifying 'interesting' residues, but cannot, at this point, assign a function unambiguously except for in rather obvious cases. It is, therefore, necessary to scan the literature associated with each high-ranking template. Note that a high score from GIRAF or SeSAW can also be used as a criterion for model selection, in cases where the alignment is ambiguous.

The inclusion of structural information enables known sequence signatures to be expended. For example, in the case of the first ATPase-like domain of RIG-I, SeSAW indicated that K30 is important. This residue turns out to be essential for RIG-I function, which now seems reasonable given its spatial proximity to the known DExH motif. Be-



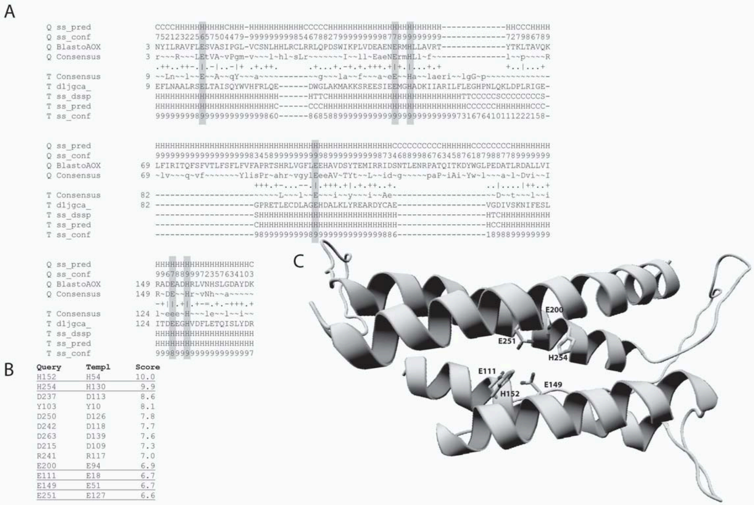
```
Q ss_pred              CEEEEECCCCHHHHHHHHHHHHHHHHHCCCCCEEEEEECCCCCCCCHHHHHHHHHHHHHCEEEEEECCCCCCCCHHHHH
Q ss_conf              948875777807899999999999999732679707887353146640258899999875225777622321577324899
Q TLR9-TIR           1 DAFVVFDKTQSAVADWVYNELRGQLEECRGRWALRLCLEERDWLPGKTLFENLWASVYGSRKTLFVLAHTDRVSGLLRAS
Q Consensus          1 DafIsYs~~D~~~~~~V~~~L~~~LE~~~~~~~~~~~l~l~~rD~~pG~~~~~~~i~~~i~~Srr~IvvlS~~~l~s~Wc~~E
                       |||||||++|   ++||.+.|.+.||    +.|+++|+++||+.||+.+.++|.++|++|+++|+|+|++|++|+||+.|
T Consensus         14 DaFISys~~D------V~~~L~~~Le------g~~~~l~~rD~~pG~~~~~~I~~~I~~Sr~~I~vlS~~~l~s~Wc~~E
T d1fyva_           14 HAFISYSGHD---SFWVKNELLPNLE----KEGMQICLHERNFVPGKSIVENIITCIEKSYKSIFVLSPNFVQSEWCHYE
T ss_dssp             EEEEECCGGG---HHHHHHTHHHHHH----TTTCCEECTTTTCCTTSCHHHHHHHHHHHEEEEEEEESHHHHHHTSHHHH
T ss_pred             EEEEEECHHH---HHHHHHHHHHHHH----HCCCEEEEECCCCCCHHHHHHHHHHHHHCCEEEEEECCHHHCCHHHHHH
T ss_conf             3688628588---9999999999998----67986998444246310268999999864788888985046158679999

Q ss_pred              HHHHHHHHHHHCCCCEEEEEECCCCCCCCC---HHHHHHHHCCCEEECCCCCCCCHHHHHHHHHHHCCCCCC
Q ss_conf              9999999987279826899755768200 25---58999873297164387667353789999986387652
Q TLR9-TIR          81 FLLAQQRLLEDRKDVVVLVILSPDGRRSRY---VRLRQRLCRQSVLLWPHQPSGQRSFWAQLGMALTRDNH
Q Consensus         81 ~~~A~~~~l~~~~~~~iI~I~le~i~~~~~~~~~~~~~l~~~L~~~t~l~W~~~~~~~~~~~~FW~~L~~al~~~~~
                       +++|+++++++++++.++|+|+++++++..++   +.|+.+++++|||+||+++.++++|||+|||+|||.+.+
T Consensus         87 ~~~A~~~~l~~~~~~~iI~V~le~i~~~~~l~~~~~~~~l~~~~~~~~yl~W~~~~~~~~~~~~FW~kLr~al~~~l~
T d1fyva_           87 LYFAHHNLFHEGSNSLILILLEPIPQYSIPSSYHKLKSLMARRTYLEWPKEKSKRGLFWANLRAAINIKLT
T ss_dssp             HHTTSCCCSCSSCSSEEEEESSCCCGGGSCTTCHHHHHHHHTCCCEECCSSGGGSSHHHHHHHHHHHSCCC
T ss_pred             HHHHHHHHHHHCCCEEEEEEECCCCHHCCCHHHHHHHHHHHCCCCEEECCCCCCCHHHHHHHHHHHCCCCCC
```

Query	Templ	Score
W136	W769	14.1
W125	W758	14.1
W16	W650	13.7
C37	C667	11.1
P45	P675	9.7
P126	P759	9.3
P103	P733	9.1
F135	F768	8.7
Y110	Y743	8.2
F3	F640	7.9
F65	F695	7.6
G46	G676	7.6
R41	R671	7.2
N52	N682	6.8
N19	N653	6.7
E26	E660	6.7
R119	R752	6.6
K62	K692	6.3
L139	L772	6.0
L25	L659	6.0
S60	S690	6.0
E40	E670	6.0
E51	E681	6.0
E20	E654	6.0
L21	L655	5.9
V17	V651	5.8
K47	K677	5.7

Figure 4: **Annotation of TLR-9 TIR domain.** The figure shows A) the raw HHpred alignment to the top conserved domain hit; B) the top-scoring residue pairs between the query and 1fyv according to SeSAW, with key residues underlined; C) the SeSAW structural model with the key residues shown as sticks.

cause the lysine in question was observed in a wide range of helicases, including one from Archaea, we can extend the DExH sequence signature to include a lysine N-terminal by approximately 100 residues. In the case of the second ATPase-like domain, a conserved helicase motif IV was found among the highest-scoring residues. By combining these two results, we can further suggest a close spatial proximity between the ATPase site and the helicase motif IV, which extends the putative sequence signature further. In the case of *Blastocystis* AOX, SeSAW located the two previously identified E. . . HxxE motifs, but also identified a number of conserved aspartic acid residues whose function is not well known. The fact that these residues are found in two proteins with different function suggests that they may serve a role in stabilizing the common fold, or support the common role of iron binding.

The agreement between SeSAW and GIRAF in the case of the first ATPase-like domain in RIG-I is encouraging. Moreover, the GIRAF p-value appears to correlate with functional significance. In the case of the TLR9 TIR domain, the lack of a GIRAF hit is consistent with its known protein-binding function, while in the one case where we are relatively confident of the ligand, AOX [11], the very low p-values validate the GIRAF statistical model.



```
Q ss_pred          CCCCHHHHHHHHHHCHHH-HHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHHHHHH-------------HHCCCHHHH
Q ss_conf          752123225657504479-999999999854678827999999999987789999999-------------727986789
Q BlastoAOX      3 NYILRAVFLESVASIPGL-VCSNLHHLRCLRRLQPDSWIKPLVDEAENERMHLLAVRT-------------YTKLTAVQK
Q Consensus      3 r~~~R~~~LEtVA~vPgm-v~~~l~hl~sLr~~~~~~~I~~ll~EaeNErmHLl~f~~~~~~~~~~~~~~~l~~p~~~~R
                   .++..++-.|.-|..-|+ .+.++.      .+...+.-+.+.+++.+|+.|...+.+             +..|...+
T Consensus      9 ~~Ln~~l~~E~~A~~qY~~~a~~~~~~~~~~~~g~~~la~~f~~~a~eE~~Ha~~laeri~~lgG~p~~~~~~~~~~~~~~
T d1jgca_      9 EFLNAALRSELTAISQYWVHFRLQE------DWGLAKMAKKSREESIEEMGHADKIIARILFLEGHPNLQKLDPLRIGE-
T ss_dssp         HHHHHHHHHHHHHHHHHHHHHHHHH------HTTCHHHHHHHHHHHHHHHHHHHHHHHHHHHTTCCCCCSCCCCCCCCS-
T ss_pred         HHHHHHHHHHHHHHHHHHHHHHHHC------CCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHCCC-
T ss_conf         9999999989999999999999860------868588999999999999999999999999731676410111222158-

Q ss_pred          HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHCCCCCCCCHHHHHHHH
Q ss_conf          99999999999999999999834589999999999899999999999987346889986763458761879887789999999
Q BlastoAOX     69 LFIRITQFSFVTLFSFLFVFAPRTSHRLVGFLEEHAVDSYTEMIRRIDSNTLENRPATQITKDYWGLPEDATLRDALLVI
Q Consensus     69 ~lv~~~q~vf~~~~~~~~YlisPr~ahr~vgylEeeAV~TYt~~L~~id~g~~~~~paP~iAi~Yw~l~~~a~l~Dvi~~I
                                       +++.+-...--.|.+++.+|.+.++..+.                  ..|....|++..+
T Consensus     82 --------------------~~~~~L~~~l~~E~~~i~~y~~~i~~Ae~~~~~~~~~~~~~~~~~~~~~D~~t~~~l~~i
T d1jgca_     82 --------------------GPRETLECDLAGEHDALKLYREARDYCAE------------------VGDIVSKNIFESL
T ss_dssp         --------------------SHHHHHHHHHHHHHHHHHHHHHHHHHHHH------------------HTCHHHHHHHHHH
T ss_pred         --------------------CHHHHHHHHHHHHHHHHHHHHHHHHHHHH------------------HCCHHHHHHHHHH
T ss_conf         --------------------98999999999989999999999999886------------------1898899999999

Q ss_pred          HHHHHHHHHHHHHHHHHHHC
Q ss_conf          99678899972357634103
Q BlastoAOX    149 RADEADHRLVNHSLGDAYDK
Q Consensus    149 R~DE~~Hr~vNh~~a~~~~~
                   -+||.+|.+-=.+.-+++.+
T Consensus    124 l~eee~H~~~le~~l~~i~~
T d1jgca_    124 ITDEEGHVDFLETQISLYDR
T ss_dssp         HHHHHHHHHHHHHHHHHHHH
T ss_pred         HHHHHHHHHHHHHHHHHHHH
T ss_conf         99989999999999999997
```

Query	Templ	Score
H152	H54	10.0
H254	H130	9.9
D237	D113	8.6
Y103	Y10	8.1
D250	D126	7.8
D242	D118	7.7
D263	D139	7.6
D215	D109	7.3
R241	R117	7.0
E200	E94	6.9
E111	E18	6.7
E149	E51	6.7
E251	E127	6.6

Figure 5: **Annotation of AOX from *Blastocystis*** The figure shows A) the raw HHpred alignment to the top conserved domain hit; B) the top-scoring residue pairs between the query and 3bkn according to SeSAW, with key residues underlined; C) the SeSAW structural model with the key residues shown as sticks.

In each case, the rank of some high-scoring residue pairs were attributed to structural conservation; such residues must be separated from those that appear to be more chemically active, and thus functionally important. When the similarity to the structural template, or the functional template, or both, is low, the location of the functional sites could be suggested but the exact function could not be determined. Such suggestions can in some cases (e.g., sulfate binding in RIG-I ATPase-like domain 1) be tested experimentally by binding assays coupled with mutational analysis. Thus the proposed method is complementary to traditional biochemical analytical methods for functional characterization.

Table 1: GIRAF results. The table shows all statistically significant ligand binding site hits from the current PDB as determined by the GIRAF p-value. In the case of the TLR-9 TIR domain, no hits were found.

Query	Pvalue	Ligand	Template	PDB ID	Comments
RIG-I ATPase-1	6x10-14	Sodium	MDA5	3b6e	Corresponds to highest scoring SeSAW site
RIG-I ATPase-2	4x10-13	sulfate	Putative transglycosylase	3czb	Significance unknown
TLR9 TIR	NA	NA	NA	NA	NA
AOX	3x10-28	Dimethyl sulfide	Designed protein	1jm0	Fe^{2+} binding site
AOX	4x10-17	Fe^{2+}	Bacterioferritin	1sof	Fe^{2+} binding site

References

[1] Standley DM, Toh H, Nakamura H. Functional annotation by sequence-weighted structure alignments: statistical analysis and case studies from the Protein 3000 structural genomics project in Japan, Proteins 2008;72:1333-1351.

[2] Kinjo AR, Nakamura H. Similarity search for local protein structures at atomic resolution by exploiting a database management system, Biophysics 2007;3:75-84.

[3] Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction, Nucleic Acids Res 2005;33:W244-248.

[4] Eswar N, Eramian D, Webb B et al. Protein structure modeling with MODELLER, Methods Mol Biol 2008;426:145-159.

[5] SeSAW. `http://pdbjs6.pdbj.org/SeSAW/` 2008.

[6] Clamp M, Cuff J, Searle SM et al. The Jalview Java alignment editor, Bioinformatics 2004;20:426-427.

[7] Kinoshita K, Nakamura H. eF-site and PDBjViewer: database and viewer for protein functional sites, Bioinformatics 2004;20:1329-1330.

[8] Yoneyama M, Fujita T. Function of RIG-I-like receptors in antiviral innate immunity, J Biol Chem 2007;282:15315-15318.

[9] Nishino T, Komori K, Tsuchiya D et al. Crystal structure and functional implications of *Pyrococcus furiosus* hef helicase domain involved in branched DNA processing, Structure 2005;13:143-153.

[10] Xu Y, Tao X, Shen B et al. Structural basis for signal transduction by the Toll/interleukin-1 receptor domains, Nature 2000;408:111-115.

[11] Affourtit C, Albury MS, Crichton PG et al. Exploring the molecular nature of alternative oxidase regulation and catalysis, FEBS Lett 2002;510:121-126.

[12] Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure, Nucleic Acids Res 2005;33:W89-93.

[13] Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments, Proc Natl Acad Sci U S A 2008;105:5441-5446.