D490–D494 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1129

Cube-DB: detection of functional divergence in human protein families

Zong Hong Zhang[1], Kavitha Bharatham[1,2], Sharon M. Q. Chee[1] and Ivana Mihalek[1,2,*]

[1]Bioinformatics Institute 30 Biopolis Street, #07-01 Matrix, Singapore 138671 and [2]School of Biological Sciences, Nanyang Technological University, 50 Nanyang Avenue, Singapore 63979

Received August 17, 2011; Revised and Accepted November 8, 2011

ABSTRACT

Cube-DB is a database of pre-evaluated results for detection of functional divergence in human/vertebrate protein families. The analysis is organized around the nomenclature associated with the human proteins, but based on all currently available vertebrate genomes. Using full genomes enables us, through a mutual-best-hit strategy, to construct comparable taxonomical samples for all paralogues under consideration. Functional specialization is scored on the residue level according to two models of behavior after divergence: heterotachy and homotachy. In the first case, the positions on the protein sequence are scored highly if they are conserved in the reference group of orthologs, and overlap poorly with the residue type choice in the paralogs groups (such positions will also be termed functional determinants). The second model additionally requires conservation within each group of paralogs (functional discriminants). The scoring functions are phylogeny independent, but sensitive to the residue type similarity. The results are presented as a table of per-residue scores, and mapped onto related structure (when available) via browser-embedded visualization tool. They can also be downloaded as a spreadsheet table, and sessions for two additional molecular visualization tools. The database interface is available at http://epsf.bmad.bii.a-star.edu.sg/cube/db/html/home.html.

INTRODUCTION

Cube-DB is designed to answer the question: which residues in a protein, belonging to a family of human paralogs, are responsible for its functional specialization? Intuitively we expect that such residues should have the same type (that is, be conserved) in related species. Whether they should be conserved as different types across paralogs in the same species has been the subject of some debate (1–5). Cube-DB takes the position that once the functional shift has occurred the conservation is no longer expected, and reports residues that are well conserved in the protein of interest and different in paralogs, irrespective of their degree of conservation.

Similar view was taken in FunShift (6). The authors of that 2005 compilation used a maximum likelihood method (7) to establish the rate of mutation across branches of a presumptive evolutionary tree. In contrast to this phylogeny-based approach, Cube-DB uses a tree-independent heuristic, to be discussed in the 'Methods' section, to estimate both within-ortholog group conservation, and the overlap (or lack thereof) across different paralogs.

SDR database (8), on the contrary, adheres to the view that the positions of functional importance should be conserved in all paralogs, an assumption that has repeatedly been shown to work well for the catalytic sites of enzymes (9,10). While Cube-DB displays this type of information side by side with the overall and group-specific conservation, it emphasizes the last characteristic—within group conservation—as a feature of practical importance in other (non-enzymatic) cases of functional divergence (11).

While several servers (that is, web applications that generate the analysis on the fly) offer specialization analysis on the set of sequences provided by the user (12–14), we choose to simplify the process by providing sets of sequences that are known to be paralogous and to align well. To do so, we limit our attention to the sets of sequences for which this information is relatively straightforward to establish, but is of preeminent interest for biomedical applications: human families of paralogous

*To whom correspondence should be addressed. Email: ivanam@bii.a-star.edu.sg
Present address:
Kavitha Bharatham, Department of Chemical Biology and Therapeutics, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, Tennessee 38105, USA

proteins and their conservation across vertebrate orthologs.

Staying within the nomenclature associated with the human versions of proteins enables us also to design a straightforward and intuitive interface for browsing the database contents. Furthermore, the database offers a unique take-home way of presenting the results, in terms of downloadable spreadsheets and sessions for two popular molecular visualization tools.

DATA PROVENANCE AND DATABASE SCOPE

We organize our analysis around the nomenclature/division into families provided by HUGO Gene Nomenclature Committee (15), but the results are equally valid for (and indeed based on) all vertebrate genomes currently available in Ensembl (16). By its design and purpose, the database is oriented toward comparison of vertebrate paralogs. Working with full genomes enables us, by using a mutual-best-hit [a.k.a. BeT, the best hit (17) or bidirectional best hit (18)] strategy, to construct relatively complete and reliable sets of orthologs from all available species, and obtain balanced sets of sequences for all paralogs under consideration. By balanced here we mean 'covering a comparable taxonomical breadth.' The question of problematic alignments is sidestepped by limiting the analysis to clusters of paralogs with at least 40% sequence similarity (19). In its current edition, Cube-DB thus presents the results for 226 named groups of paralogs, divided into 600 clusters of alignable sequences.

METHOD

Assembling and aligning the relevant sequence set

Cube-DB subdivides the list of human protein families provided by HUGO (13) into clusters of proteins with at least 40% sequence identity in at least 70% of their alignable (non-gap) length. The purpose is 2-fold: it eliminates the problem of ambiguous alignments (19), and it helps divide the results into tractable chunks for presentation. Different choices (and sizes) of groups to be compared are, of course, possible, and should at some point be available through the accompanying server.

For human sequences belonging to a cluster, the orthologs from other vertebrate species from Ensembl (16) are retrieved by mutual-best-hit strategy. For a recent comparison of mutual-best-hit approach with other available options, see (20). The taxonomical content of the database is thus entirely determined by the vertebrate genomes currently (Release 64) deposited in Ensembl.

When an ortholog is reported to be missing in the database of known (annotated) proteins from a genome, it is sought in the *ab initio* detected set of proteins.

Each set of orthologs is aligned using Mafft (21) and the resulting alignments (corresponding to a single human paralog each) are then profile-aligned using the same program.

Assigning relative conservation and specialization scores to each position in the alignment

The algebraic expressions used to evaluate the scores described below can be found in the Supplementary Data. An extensive discussion can be found in (11).

For each position in the overall alignment the conservation is scored on the [0, 1] scale. Similarly, for each pair of paralogous groups, the overlap in the amino acid type choice is turned into a quantity in the same range. In addition, the scoring functions are sensitive to similarity between the amino acid types—both conservation and overlap are measured from their expected values given the distribution of the amino acid types at a position, the overall variability in the alignment, and the average propensity of residue types to mutate into each other [see Supplementatry Data and (11)].

These elementary scores are linearly combined into two different kinds of specialization scores, rewarding: (i) discriminants—positions that are conserved in each paralog as a different amino acid type, and (ii) determinants—positions that, referring to a particular paralog, are conserved in that group, and different in the non-reference groups, irrespective of their conservation therein.

For comparison, the overall conservation across all sequences is calculated using a previously published method (16). Highly conserved positions do not coincide with the specific positions, and correspond to structural and functional features common to all paralogs in the cluster.

Database organization

Starting from the list of family names, which is small from a computational perspective, and clustering the paralogs by similarity, results in a shallow directory structure that can be quickly traversed without using a database management software. All the result files corresponding to a cluster specified in the query are located therein directly, and returned to the user.

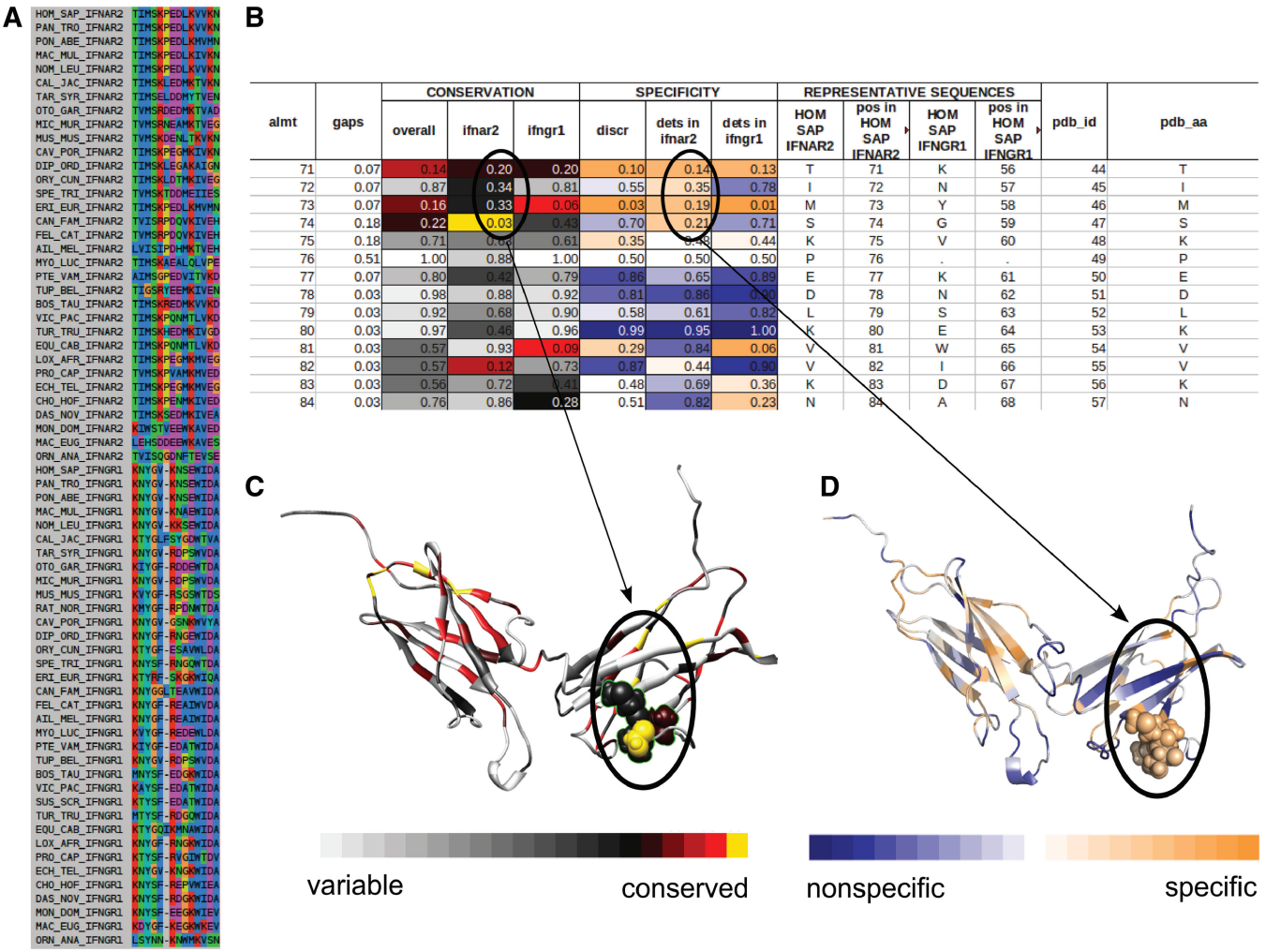
RESULTS AND THEIR PRESENTATION

Input

The database can be browsed though alphabetically ordered HUGO nomenclature, or searched by protein name, using a simple string matching search.

Output

Per-residue results of estimated degree of conservation across families, as well as two different models of functional divergence [discriminants and determinants; also termed type I and type II (1), heterotachy and homotachy (2) in the molecular evolution literature] are presented in terms of a scrollable html table, and embedded Jmol (22) visualization tool, for the cases when the related structure is available. These results are also available for download in terms of an xls spreadsheet table, and sessions for two different protein visualization tools: Pymol (23) and Chimera (24), Figure 1. To keep the size of visualization sessions manageable, visualization for determinants of each paralogous group is presented as an individual



almt	gaps	CONSERVATION overall	ifnar2	ifngr1	SPECIFICITY discr	dets in ifnar2	dets in ifngr1	REPRESENTATIVE SEQUENCES HOM SAP IFNAR2	pos in HOM SAP IFNAR2	HOM SAP IFNGR1	pos in HOM SAP IFNGR1	pdb_id	pdb_aa
71	0.07	0.14	0.20	0.20	0.10	0.14	0.13	T	71	K	56	44	T
72	0.07	0.87	0.34	0.81	0.55	0.35	0.78	I	72	N	57	45	I
73	0.07	0.16	0.33	0.06	0.03	0.19	0.01	M	73	Y	58	46	M
74	0.18	0.22	0.03	0.43	0.70	0.21	0.71	S	74	G	59	47	S
75	0.18	0.71	0.66	0.61	0.35	0.48	0.44	K	75	V	60	48	K
76	0.51	1.00	0.88	1.00	0.50	0.50	0.50	P	76	.	.	49	P
77	0.07	0.80	0.42	0.79	0.86	0.65	0.89	E	77	K	61	50	E
78	0.03	0.98	0.88	0.92	0.81	0.86	0.90	D	78	N	62	51	D
79	0.03	0.92	0.68	0.90	0.58	0.61	0.82	L	79	S	63	52	L
80	0.03	0.97	0.46	0.96	0.99	0.95	1.00	K	80	E	64	53	K
81	0.03	0.57	0.93	0.09	0.29	0.84	0.06	V	81	W	65	54	V
82	0.03	0.57	0.12	0.73	0.87	0.44	0.90	V	82	I	66	55	V
83	0.03	0.56	0.72	0.41	0.48	0.69	0.36	K	83	D	67	56	K
84	0.03	0.76	0.86	0.28	0.51	0.82	0.23	N	84	A	68	57	N

Figure 1. Result presentation in Cube-DB. (**A**) Several columns from the alignment of IFNAR2 and IFNGR1, members of interferon receptor family. (**B**) The region from downloadable spreadsheet, corresponding to the same region as shown in (A), collating the information about conservation (white–black–red colorbar) with the information about specialization (blue–orange colorbar). The ''conservation'' group of of columns shows conservation across all groups, as well as within each group of orthologs in the cluster. The 'specificity' group of column shows the scoring of discriminant behavior, which is a property of the cluster as a whole (see 'Methods' section), and determinant behavior, using each group as a reference in turn. (**C**) Visualization of conservation in IFNAR2, using downloadable Chimera (24) session. (**D**) Visualization of specificity determinants in IFNAR2, using downloadable Pymol (23) session. The information about the same group of residues is encircled in the table and on the visualization frame, to illustrate the correspondence of the color coding between the two.

session. The alignments used in the analysis, as well as complete work directory are also available for download to interested users. Help pages are accessible from each page presented by the web.

User's perspective: an example. We illustrate the database functionality on the example of a hypothetical user investigating the sources of functional difference between interferon-α receptor 2 (IFNAR2) and its cousin, interferon-γ receptor 1 (IFNGR1). The reason we choose this example is that the thorough mutational study undertaken by Piehler and Schreiber (25) enables us to take a look at the results presented by the database in the retrospective light.

The related analysis can be found by locating the IFNR (Interferon receptors) family on the Browse page, followed by narrowing the search down to 'cluster_2', a subgroup of the family consisting of two members: IFNAR2 and IFNGR1. The same page can be located through the Search window, using 'IFNAR2' or 'IFNGR1' as search terms.

Since for our protein of interest the structure is available, the top of the results page shows side-by-side visualization of overall conservation and discriminant specialization scores, mapped onto the structure. Discriminant behavior here refers to positions that are conserved within a group of orthologs, but as a different residue type within each group. To keep the visual cue as clear as possible, we choose to use two different coloring schemes for the two properties (conservation and specialization). The colorbars shown in Figure 1 also appear on the top of the results page. The color scales we choose - mainly because they have very little overlap—are white–black–red for conservation and blue–orange for

specialization. This 3D mapping of the results brings to the attention potential clustering of residues on one face of the protein, that otherwise might go unnoticed on the sequence.

The same information, using the same color-coding, can be found in the table below the visualization windows. Therein the numerical value of the scores is also given. Additionally, the table columns contain the information about the within-group conservation and specialization scoring according to determinant model. This last type of scoring is reference group-dependent (see 'Methods' section), and therefore one 'determinant' column for each of the groups appears in the table.

However, this kind of result presentation is limited by the browser's capabilities, and is, furthermore, inflexible and unmodifiable. A typical user will already have a sizable knowledge (public or proprietary) about the protein under investigation, that can extend and be compared with the conservation/specialization analysis. For that purpose the Downloads page for each cluster contains a number of downloadable files for use in a spreadsheet application or molecular visualization tools. Spreadsheet allows for a further modification of the results' presentation: the residues can be sorted according to any of the provided scores, as needed by the researcher.

These files are further divided according to the underlying selection of sequences, which can cover all available vertebrates, or mammalian sequences only. They can be extended with additional information, and saved to be used as a reference.

Figure 2 shows the correspondence between the mutational data of Piehler and Schreiber (25), and the specialization scoring using mammalian sequences and determinant model of specialization (see 'Methods' section above), applied to IFNAR2. The upper panel of the figure shows as spheres the positions [see (11) and its supplementary material] found to be involved in the function specific for IFNAR2 (binding of interferon α 2, IFNα2 and interferon β, IFNβ), which are mostly scored favorably (orange) by the scoring method. The lower panel shows the positions without functional impact on the binding between IFNAR2 and IFNα/β. They are scored unfavorably in the same scoring scheme.

In a real-life scenario, the degree of involvement of individual residues in functional specialization is unknown, and is indeed the object of the study. Focusing on residues of strong specialization (orange) should help locate candidate regions of group-specific functional impact. Such residues will typically be found interspersed with conserved residues in the ordered pieces of secondary protein structure, or forming larger continuous stretches in disordered regions.

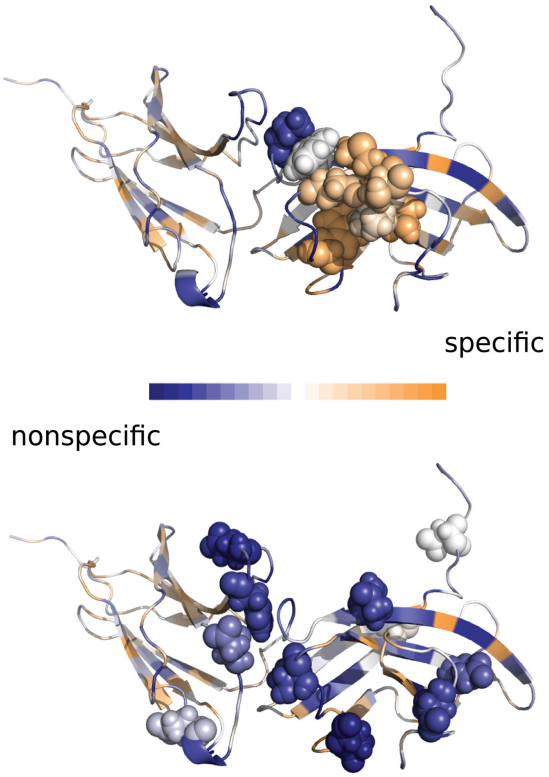


Figure 2. Comparison of determinant scoring for IFNAR2 and the results of Piehler and Schreiber (25). The results are mapped on the structure of IFNAR2 determined by Nudelman *et al.* (26) (PDB identifier 2lag). The positions tested in the experiment are shown as spheres. The coloring scheme corresponds to determinant model scoring, as implemented in Cube-DB, and applied to mamalian orthologues of the two proteins. **(A)** The positions shown in experiment [(25), Table 2.] to be involved in a function specific to IFNAR2 (binding of interferon α 2, IFNα2 and interferon β, IFNβ). **(B)** The positions shown in experiment *not* to be involved in a function specific to IFNAR2.

CONCLUSION AND OUTLOOK

Cube-DB offers a unique service for detecting residues responsible for functional specialization in human protein families. Its largest value lies in collating several scores—for conservation within and across several groups of paralogs, as well as divergence between them—and presenting them in a form that can be downloaded and extended with further annotation by the user. Beyond doubt, the result presentation can be elaborated and improved on in several ways, for example by linking dynamically the (so far physically independent) visualizations for the alignments, the scores, and their mapping onto the structure. We hope to return to this possibility in one of the subsequent versions of the database. Also, the limitation to the protein families currently recognized by the community consensus will be amended in the future by case-specific extensions of the database, and through the (planned) accompanying server.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary References [1–3].

ACKNOWLEDGEMENTS

Funding by Biomedical Research Council of A*STAR Singapore is gratefully acknowledged.

FUNDING

Funding for open access charge: Agency for Science Technology and Research, Singapore.

Conflict of interest statement. None declared.

REFERENCES

1. Gu,X. (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.*, **18**, 453–464.
2. Lopez,P., Casane,D. and Philippe,H. (2002) Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.*, **19**, 1–7.
3. Capra,J. and Singh,M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, **24**, 1473–1480.
4. Lichtarge,O., Yamamoto,K. and Cohen,F. (1997) Identification of functional surfaces of the zinc binding domains of intracellular receptors1. *J. Mol. Biol.*, **274**, 325–337.
5. Madabushi,S., Gross,A., Philippi,A., Meng,E., Wensel,T. and Lichtarge,O. (2004) Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *J. Biol. Chem.*, **279**, 8126–8132.
6. Abhiman,S. and Sonnhammer,E. (2005) FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.*, **33**, D197–D200.
7. Knudsen,B. and Miyamoto,M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl Acad. Sci. USA*, **98**, 14512–14517.
8. Donald,J. and Shakhnovich,E. (2009) SDR: a database of predicted specificity-determining residues in proteins. *Nucleic Acids Res.*, **37**, D191–D194.
9. Chakrabarti,S. and Panchenko,A. (2009) Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics*, **10**, 207.
10. Rausell,A., Juan,D., Pazos,F. and Valencia,A. (2010) Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc. Natl Acad. Sci. USA*, **107**, 1995–2000.
11. Bharatham,K., Zhang,Z.H. and Mihalek,I. (2011) Determinants, discriminants, conserved residues - a heuristic approach to detection of functional divergence in protein families. *PLoS One*, **6**, e24382.
12. Kalinina,O., Mironov,A., Gelfand,M. and Rakhmaninova,A. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, **13**, 443–456.
13. Carro,A., Tress,M., De Juan,D., Pazos,F., Lopez-Romero,P., Del Sol,A., Valencia,A. and Rojas,A. (2006) TreeDet: a web server to explore sequence space. *Nucleic Acids Res.*, **34**, W110–W115.
14. Chakrabarti,S., Bryant,S. and Panchenko,A. (2007) Functional specificity lies within the properties and evolutionary changes of amino acids. *J. Mol. Biol.*, **373**, 801–810.
15. Seal,R., Gordon,S., Lush,M., Wright,M. and Bruford,E. (2011) genenames. org: the HGNC resources in 2011. *Nucleic Acids Res.*, **39**, D514–D519.
16. Flicek,P., Amode,M., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al*. (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
17. Tatusov,R., Koonin,E. and Lipman,D. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
18. Overbeek,R., Fonstein,M., DSouza,M., Pusch,G. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
19. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*, **12**, 85–94.
20. Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.
21. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33**, 511–518.
22. *Jmol; An open-source Java viewer for chemical structures in 3D*. http://www.jmol.org (21 November 2011, date last accessed).
23. DeLano,W. *The PyMOL Molecular Graphics System*. http://www.pymol.org (21 November 2011, date last accessed).
24. Pettersen,E., Goddard,T., Huang,C., Couch,G., Greenblatt,D., Meng,E. and Ferrin,T. (2004) UCSF Chimera visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
25. Piehler,J. and Schreiber,G. (1999) Mutational and structural analysis of the binding interface between type I interferons and their receptor ifnar2. *J. Mol. Biol.*, **294**, 223–237.
26. Nudelman,I., Akabayov,S., Scherf,T. and Anglister,J. (2011) Observation of intermolecular interactions in large protein complexes by 2D-double difference NOESY: application to the 44 kDa interferon-receptor complex. *J. Am. Chem. Soc.*, **133**, 14755–14764.