

№ 15

Phyre and PhyreRisk – Integrating genetic variant data with experimental and predicted protein structures and their complexes

Lawrence A Kelley, Alessia David, Sirawit Ittisoponpisan, Stefans Mezulis, Tochukwu C Ofoegbu, Devlina Chakravarty, Petras J Kundrotas, Ilya A Vakser and Michael Sternberg

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

March 16, 2018

Phyre and PhyreRisk – Integrating genetic variant data with experimental and predicted protein structures and their complexes

Lawrence A Kelley (1), Alessia David (1), Sirawit Ittisoponpisan (1), Stefans Mezulis (1), Tochukwu C Ofoegbu (1), Devlina Chakravarty (2), Petras J Kundrotas (2), Ilya A Vakser (2), and Michael J E Sternberg (1)

- (1) Structural Bioinformatics Group, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK
- (2) Computational Biology Program and Center for Computational Biology, Department of Molecular Biosciences, The University of Kansas, 2030 Becker Drive, Lawrence, KS 66045 USA

Abstract

Phyre is a widely-used web server for tertiary protein structure prediction and modelling with over 80,000 distinct users per year and is part of the ELIXIR tools resource. We are developing 3D models for predicted complexes by integrating tertiary models, the experimental structure of complexes and protein-protein interaction data. A new resource, PhyreRisk, is being developed which will enable users to map protein missense variants onto experimental and predicted tertiary structures and complexes. See <u>www.sbg.bio.ic.ac.uk/~phyre2</u> and <u>phyrerisk.bc.ic.ac.uk</u>

1. Introduction

Knowledge of the 3D structure of protein and their complexes is central to understanding biological behaviour and to the development of strategies to exploit this fundamental knowledge. In particular, the rapid decreasing cost of genome and exome sequencing is yielding a wealth of data about genetic variation including missense (non-synonymous) variants. The prediction the probable phenotypic effect of such variants, particular if they are associated with a diseased state, can be markedly enhanced via structural information. We report PhyreRisk which integrates variant and structural data.

2. Phyre2 and Phyre3

Phyre2 (1) provides a web-based resource for the community to obtain predicted 3D structures based on the known structure of a template. Predicted models typically are returned via an e-mail link within a few hours and the user can examine alternate models with their multiple sequence alignments. There are facilities for users to examine the accuracy of the model along the sequence (Phyre Investigator), to request weekly re-runs of their query against updates databases (PhyreAlarm), and to see if a 3D structure occurs in other genomes (BackPhyre). Users can also submit batch files with numerous sequences to process. Phyre2 is now a resource with the UK node of ELIXIR, a European distributed infrastructure for life-science information. There are over 80,000 distinct users of Phyre2 per year. We are launching Phyre3 which will have an enhanced interface with more information in different, toggable windows.

3. Prediction of complexes

With the increase in the number of experimentally-determined structures of complexes, a powerful and widely applicable approach is to use these as templates to predict the structure of other complexes. The Kansas group has been following this approach and based on identified protein-protein interactions in Intact and BioGrid has provided the community with the GWIDD (2) database of predicted complexes. We are now providing the Kansas group with predicted Phyre models, which together with experimental structures, are the inputs for the prediction of binary complexes.

4. PhyreRisk

PhyreRisk, which we plan to release in summer 2018, will provide a resource for a user with human variant data, either at the genomic or the protein level, to map any missense, nonsense or stop gain variant onto the structure of an individual protein and its complex. PhyreRisk has an interactive sequence/structure viewer to provide user-driven mapping. Moreover we have developed an approach to predict the structural consequences of a missense mutation (such as loss of a salt bridge) and have parameterised this to be effective in both experimental and predicted structures. Our plan is to integrate databases of variation, such as ClinVar and gnomAD, into PhyreRisk.

	PHYRERIS	SK		Home About	Contact Us Help+	Sign in		
O94929			c	U	3D Structu	are Viewer - JSMol		Ξ
	Sequ	ence Browser	1					
1 MNTSIPYQQNPYNPRGS								
101 YHPKCFVCSLCRKPFP1 101 DYHAQFGIKCETCDRY1 101 NETLNYKDLAALPKVKS	SNVIQCYRCGDTCKGEVVRVHNNHFHIRCF IGDKVTFSGKECVCQTCSQSMASSKFIKIRG SGRVLEAGGKHYHFTCARCVRCHQMFTBGE SITYEVQRPDLISYEPHSRYKSDEMLERCCYG IRSSTPTSYQAPKHFHIPAGDSNIYRKPFIY	PSHCAGCKEEIKHGUSLLALDKUWHVSC EMYLTGSEVWHPICKQAARAEKKLKHRR ESLGTLSPYSODIVENLDLRORBASSPG	PREQUESVILTGEVISKDGVPYCES 200 ISETSISPPGSSIGSPNRVICAKVD 300 VIDSPTYSROGMSPTFSRSPHHYVR 400					
	RSSTPTSYQAPKHFHIPAGDSNIYRKPPIY KLOSGIGRLILKEENKARSSSYADPWTPPRS	KRHGDLSTATŘSKTSEDISOŤSKYSPIY STSSREALHTAGYEMSLNGŠPRSHYLAD	SPDPYYASESEYWTYHGSPKVPRAR 500 SDPLISKSASLPAYRRNGLHRTPSA 600					
Residue Residue In Res		HESOBERYOUNGMIT SEEDREADWKRNE	TROARDE 700		1			
Tooltip	ILA (610)				2	1000		
					~			
						~	5m	
Residue: ALA (610)							42	
	Available struct	ures (Graph-View Mode)	-	Correspoir	Iding	2P		
				position in Structure		00	\searrow	
 UniProt (Ref Seq) 		 Displayed Structure 					-)	
 Experimental (XRa) 		Experimental (NMR)					-	
	A 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1							
 Modelled (Phyre) 	 CT000 		51	3				
	• · · · · · · ·		61	3				
			61	3				
			6	BDB - tuir	Mathod - N	1140	Paralistian : 10	JSmo
	umma		c.	PDB : 1ujs Length : 74	Method : N Coverage :		Resolution : 10 Chain(s) : A	
	umma	ts (From UniProt)	6		Coverage :			
Modelled (Phyre)	Varian	ts (From UniProt)	Position	Length: 74	Coverage : other info goes here	609 - 683	Chain(s) : A	
Modelled (Phyre)	Varian Variation Missing	ts (From UniProt)	1 - 514	Length: 74	Coverage : other info goes here		Chain(s) : A	
Modelled (Phyre)	Varian Varian Missing Missing	ts (From UniProt)	1 - 514 402 - 450	Length: 74	Coverage : other info goes here Protein Ir	609 - 683 nfo (From UniProt)	Chain(s) : A	
Modelled (Phyre) Type splice variant splice variant splice variant	Varian Variation Missing Missing	ts (From UniProt)	1 - 514 402 - 450 402 - 434	Length: 74	Coverage : other info goes here Protein Ir	609 - 683	Chain(s) : A	
Modelled (Phyre)	Varian Variation Missing Missing G → D	ts (From UniProt)	1 - 514 402 - 450 402 - 434 125	Length : 74 Other Information: 6	Coverage : other info goes here Protein In Isoform	609 - 683 nfo (From UniProt)	Chain(s) : A	
Modelled (Phyre)	Varian Variation Missing Missing G + D Missing		1 - 514 402 - 450 402 - 434 125 297 - 358	Length : 74 Other information: o	Coverage : ather info goes here Protein Is Isoform Remark	609 - 683 nfo (From UniProt)	Chain(s) : A	
Modelled (Phyre)	Varian Variation Missing Missing G + D Missing	ts (From UniProt) MTISEFDRLALWKRNELKKQARLF →	1 - 514 402 - 450 402 - 434 125	Length : 74 Other Information: 6	Coverage : other info goes here Protein Ir Isoform Remark ed)	609 - 683 nfo (From UniProt)	Chain(s) : A	
Modelled (Phyre)	Varian Variation Missing Missing G → D Missing RHSQEEProyPG		1 - 514 402 - 450 402 - 434 125 297 - 358	Length : 74 Other Information: 4 Id (Type) 094929-1 (display 094929-2 (describ	Coverage : ather info goes here Protein In Isoform Remark ed) No experir available.	609 - 683 nfo (From UniProt) Is (From UniProt)	Chain(s) : A Ref VSP_012127 VSP_012130	0.0 Å VSP_012128
Modelled (Phyre)	Varian Variation Missing Missing G + D Missing RNLSQEEP(QVFG GNVWGGCL		1 - 514 402 - 450 402 - 434 125 297 - 358 647 - 683	Length : 74 Other information: 4 id (Type) 094929-1 (display	Coverage : sther info goes here Protein In Isoform Remark ed) No experir available. ed) No experir	609 - 683 nfo (From UniProt) Is (From UniProt)	Chain(s) : A Ref VSP_012127	0.0 Å VSP_012128
Modelled (Phyre)	Varian Variation Missing G → D Missing RHLSQEEFYQVFG GNPWKSGCL K → R		1 - 514 402 - 450 402 - 434 125 297 - 358 647 - 683	Length : 74 Other Information: 4 Id (Type) 094929-1 (display 094929-2 (describ	Coverage : Coverage : Protein In Isoform Remark ed) No experi available. ed) No experi available.	609 - 683 nfo (From UniProt) is (From UniProt) mental confirmation	Chain(s) : A Ref VSP_012127 VSP_012130	0.0 Å VSP_012128
Modelled (Phyre)	Varian Variation Missing Missing G → D Missing RetLSQEErQvPG GNPWGGCL K → R Interacti	MTISEFORLALWKRNELJKKQARLF →	1 - 514 402 - 450 402 - 434 125 297 - 558 647 - 683 227	Length : 74 Other Information: 0 Id (Type) 094929-1 (display 094929-2 (describ	Coverage : Coverage : Protein In Isoform Remark ed) No experi available. ed) No experi available.	609 - 683 fo (From UniProt) s (From UniProt) mental confirmation	Chain(s) : A Ref VSP_012127 VSP_012130 VSP_012127 VSP	0.0 Å VSP_012128
Modelled (Phyre)	Varian Variation Missing Missing G → D Missing RH.5gcEFroyPG GAPPAKSGCL K → R Interacti Entry	MTISEFORLALWRRNELAKQARLF → Ions (From UniProt) IntAct	1 - 514 402 - 450 402 - 434 125 297 - 358 647 - 683 227 Exp	Length : 74 Other Information: of Id (Type) 094929-1 (display 094929-2 (describ	Coverage : coverage : Protein In Isoform Remark ed) No experir available. ed) No experir available.	609 - 683 fo (From UniProt) s (From UniProt) mental confirmation	Chain(s) : A Ref VSP_012127 VSP_012130 VSP_012127 VSP	0.0 Å VSP_012128
Modelled (Phyre)	Varian Variation Missing Missing G → D Missing RetLSQEErQvPG GNPWGGCL K → R Interacti	MTISEFORLALWKRNELJKKQARLF →	1 - 514 402 - 450 402 - 434 125 297 - 558 647 - 683 227	Length : 74 Other Information: of Id (Type) 094929-1 (display 094929-2 (describ	Coverage : coverage : Protein In Isoform Remark ed) No experir available. ed) No experir available.	609 - 683 fo (From UniProt) s (From UniProt) mental confirmation	Chain(s) : A Ref VSP_012127 VSP_012130 VSP_012127 VSP	0.0 Å VSP_012128

A screenshot of PhyreRisk (taken 5 March 2018). The display show which experimental and Phyrepredicted structures that map to the sequence. Selection of one of these structures generates the molecular graphics display. There is an interactive mapping of a residue from the sequence onto the structure.

5. Discussion

One challenge we address is to develop a method to present users with a representative subset of predicted structures (e.g. open and closed forms) whilst not overburdening them with trivial different structures (which can be inspected if required). We aim to provide a measure of confidence given that complexes can be based on one or two predicted tertiary structures and then modelled into a predicted model for their association. Despite major advances in linking database, as we work on these resources we are need to resolve many technical issues to deliver a resource that meets the needs of the ever-increasing number of users who are not bioinformaticians. This work was funded by the BBSRC (BB/M011526/1 & BB/P011705/1), the Wellcome Trust (104955/Z/14/Z), the NSF and the Thai Government (SI).

References

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, *10*(6), 845. Kundrotas, P. J., Zhu, Z., & Vakser, I. A. (2012). GWIDD: a comprehensive resource for genome-wide structural modeling of protein-protein interactions. *Human genomics*, *6*(1), 7.