



Hartree Centre

Science & Technology Facilities Council

Enabling TDM Tool Development

Dr Rob Firth

Senior AI Research Scientist, STFC Hartree Centre

Text and Data Mining: what is real and applicable?

Panel @ STM Innovations 3/12/2019

The Hartree Centre



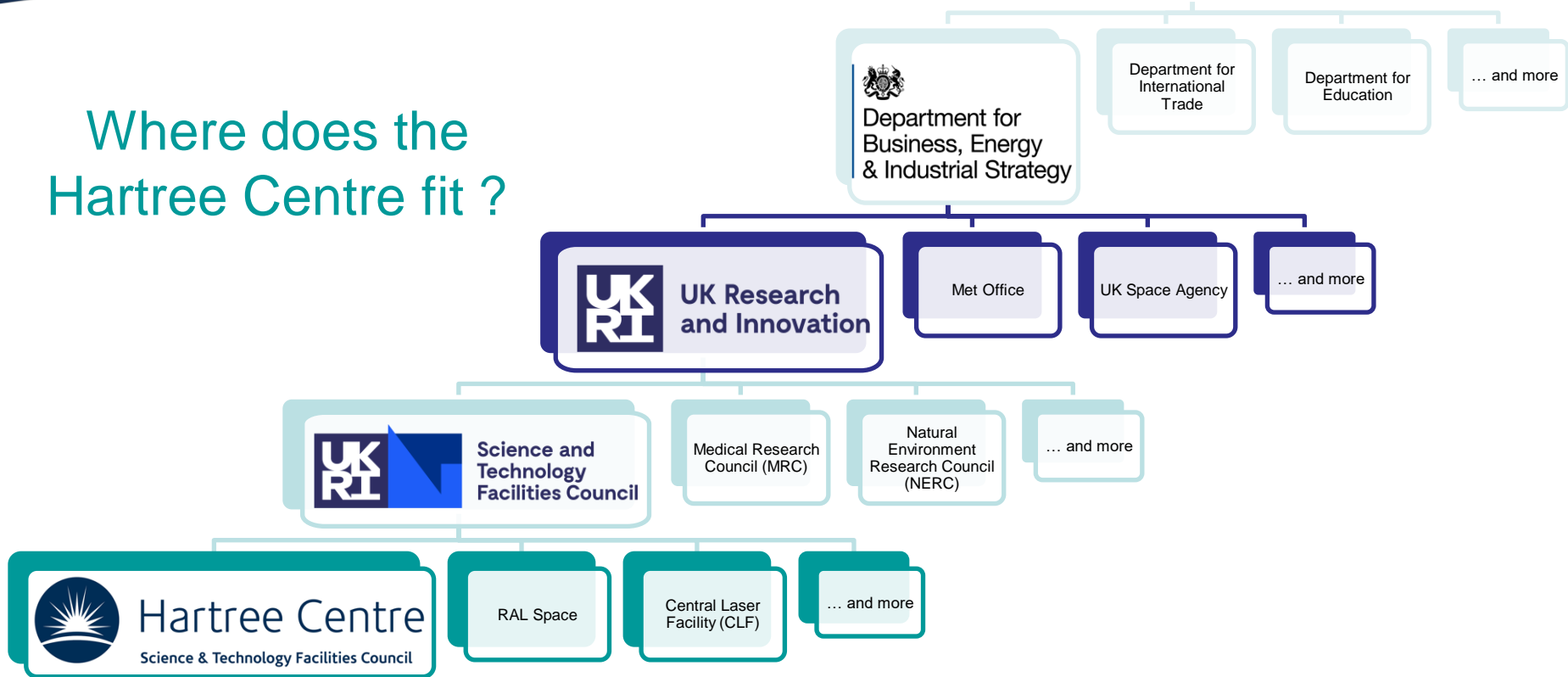
Located at the STFC Daresbury Laboratory on the Sci-Tech Daresbury campus.

Mission: To transform UK industry by accelerating the adoption of high performance computing, data-centric computing and AI.

Staff based in the building include:

- STFC
- IBM Research
- University of Liverpool Virtual Engineering Centre

Where does the Hartree Centre fit ?



What we do

– Collaborative R&D

Address industrial, societal and scientific challenges.
Bespoke small teams built around challenge-led industry projects.

– Platform as a service

Give your own experts pay-as-you-go access to our compute power

– Creating digital assets

License the new industry-led software applications we create with IBM Research

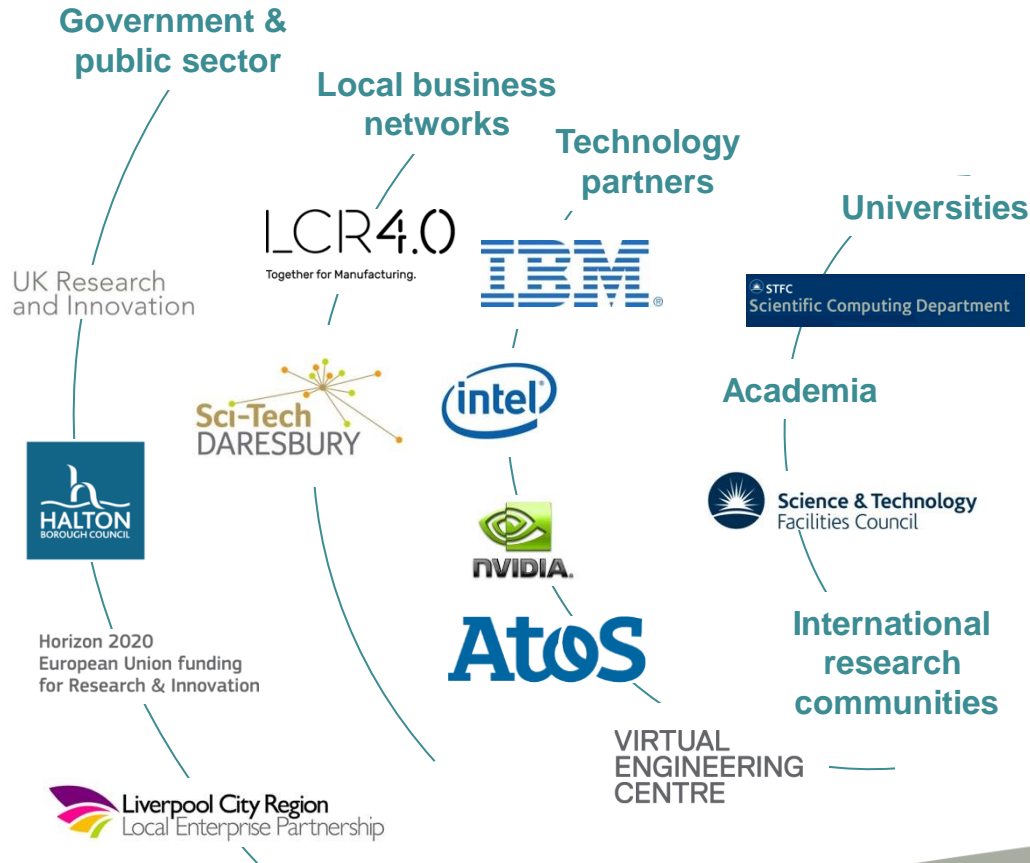
– Training and skills

Drop in on our comprehensive programme of specialist training courses and events or design a bespoke course for your team.

- 60+ computational scientists and technologists
- Insights into new & evolving technologies e.g. HPC, AI, Quantum Computing.



Our Network



Hartree Centre

Science & Technology Facilities Council

Our Network



Collaborators



Jo McEntyre



Sameer
Velankar



Hartree Centre
Science & Technology Facilities Council



Chris Morris



Rob Firth



Aravind
Venkatesan



Francesco
Talo



Abhik
Mukhopadhyay



Hartree Centre
Science & Technology Facilities Council

Collaborators



Jo McEntyre



Sameer Velankar



Hartree Centre
Science & Technology Facilities Council



Chris Morris



Rob Firth



Aravind Venkatesan



Francesco Talo



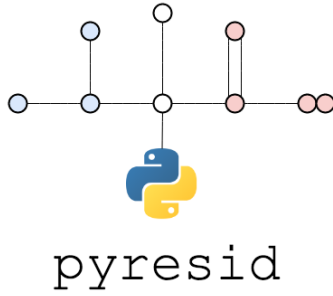
Abhik Mukhopadhyay



Hartree Centre
Science & Technology Facilities Council

TDM - Real and Applicable?





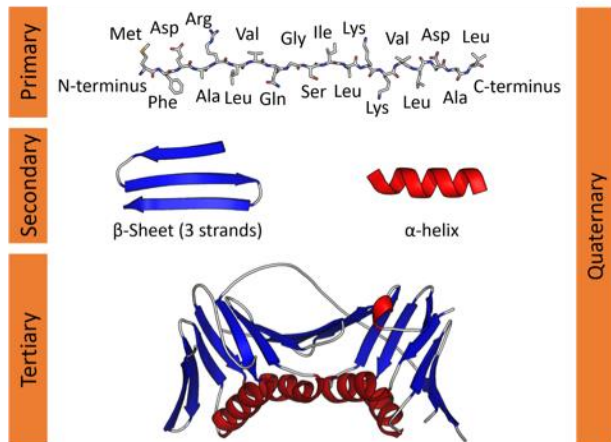
Protein Residue Annotator

Case Study

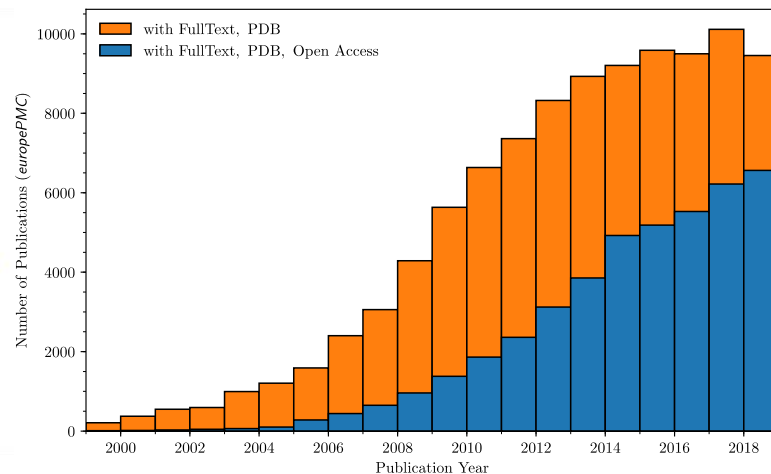


Hartree Centre
Science & Technology Facilities Council

Background: Proteins, Structural Biology



Credit: Thomas Shafee, see: [Protein Structure, Wikipedia](https://en.wikipedia.org/wiki/Protein_Structure)



Credit: R. Firth

- Proteins can be large, with only a small site of interest for a researcher/experiment
- Small primary-structural changes can cause large effects on macro-structure and interaction behavior
- Ever mounting literature to search (just like everywhere else!)
 - Speed up search
 - Augment the corpus



Background: Proteins, Structural Biology

Fortunately:

References to residues in text are (usually) referred to in a well constrained way:

Residue Name followed by **Position Number**.

Tyr33, His-455, Lys(382), Methionine 120
Serine at position 91, leucine residue at position 16, F123 residue

Glu30-Tyr33-Trp71, Ala29-33
Trp71/77, Cys99/Pro101, Arg(96)/(102), Ser-91/Gly-97

Cys residues at positions 73 and 152,
Arginine at position 452, 459 and 466



Hartree Centre

Science & Technology Facilities Council

Background: Proteins, Structural Biology

Fortunately:

References to residues in text are (usually) referred to in a well constrained way:

Residue Name followed by **Position Number**.

Tyr33, His-455, Lys(382), Methionine 120
Serine at position 91, leucine residue at position 16, F123 residue

Glu30-Tyr33-Trp71, Ala29-33
Trp71/77, Cys99/Pro101, Arg(96)/(102), Ser-91/Gly-97

Cys residues at positions 73 and 152,
Arginine at position 452, 459 and 466

Also Fortunately:

- Relevant bioinformatics resources are well established and maintained
- Recent platform upgrades allow good API access and linking across resources
- Recent development on **Natural Language Processing** (NLP) pipelines allow at-scale performance off-the-shelf
- Common Structural data format:
PDBx/mmCIF

Source:

Text, Tables, Metadata



Bioinformatics APIs



NLP Pipelines



Hartree Centre

Science & Technology Facilities Council

Automatic Annotation of Protein Residues in Published Papers

methods communications



ISSN 2053-230X

Received 6 December 2018
Accepted 1 September 2019

Edited by J. Newman, Bio21 Collaborative
Crystallisation Centre, Australia

Keywords: NLP; natural language processing;
named entity recognizer; residue.

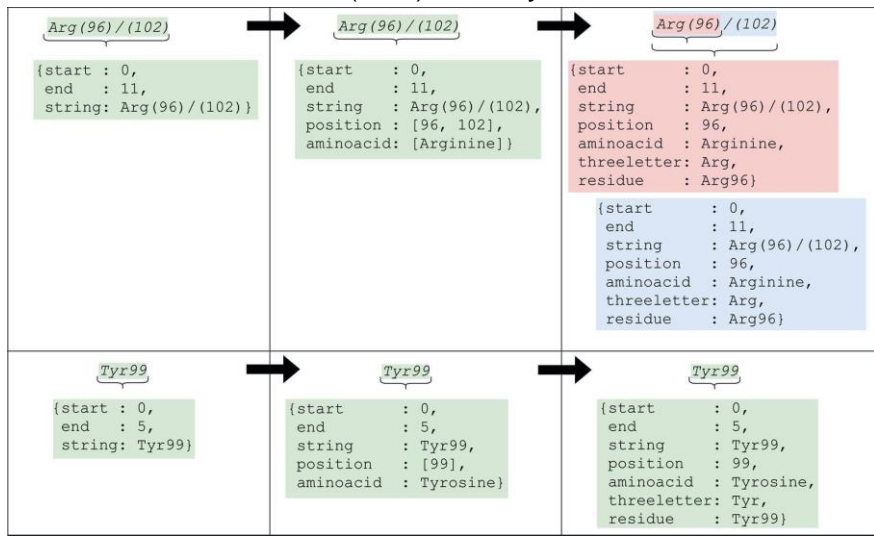
Automatic annotation of protein residues in published papers

Robert Firth,^{a*} Francesco Talo,^b Aravind Venkatesan,^b Abhik Mukhopadhyay,^b Johanna McEntyre,^b Sameer Velankar^b and Chris Morris^a

^aSTFC, Daresbury Laboratory, Warrington WA4 4AD, England, and ^bEuropean Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, England. *Correspondence e-mail: robert.firth@stfc.ac.uk

This work presents an annotation tool that automatically locates mentions of particular amino-acid residues in published papers and identifies the protein concerned. These matches can be provided in context or in a searchable format in order for researchers to better use the existing and future literature.



Firth, R., Talo, F., Venkatesan, A., Mukhopadhyay, A., McEntyre, J., Velankar, S. & Morris, C. (2019). *Acta Cryst. F* **75**, 665–672.



1. Ingest source text

 Europe PMC, or user provided

2. Identify Residues and Proteins

 PDBe URI,  UniProt URI
Protein Data Bank in Europe

3. Associate Residues with Proteins

PDBx/mmCIF from  PDBe
Protein Data Bank in Europe

4. Output Annotations

 Europe PMC SciLite, JSON



Hartree Centre
Science & Technology Facilities Council



- Annotations produced by Pyresid appear alongside a suite of other ‘BioEntities’:
 - e.g. diseases, chemicals, protein interactions
- Platform is ePMC “SciLite” (Venkatesan et al. 2017; <https://europepmc.org/Annotations>)
- Highlighted and searchable
- Links to Uniprot entries on click
- Currently between Pyresid versions



Figure 3

Ribbon representation of Ocj structures with octopine in pink/magenta for the arginine/pyruvate part, respectively. (a) Lobes 1 and 2 are shown in cyan and orange, respectively, and the hinge region in red. (b) Comparison between the open unliganded ...

Structural comparison between Ocj-octopine and NocT-octopine complexes: a different octopine binding mode

The octopine bound between the two closed lobes of Ocj is very well defined in the electron density maps (Fig. 3c), and is surrounded by 18 residues defining the ligand binding site of Ocj (Table 3). Both structures of Ocj and NocT in complex with octopine (PDB code 5ITP for NocT-octopine,¹⁴) superimpose with an average RMSD of 1.7 Å over all Ca atoms. They share a very similar binding site around the arginine moiety of octopine (Table 3 and Fig. 3c,d). Indeed, the guanidyl side chain of arginine is wedged between two conserved aromatic residues (Tyr33/39 and Trp71/77 in Ocj/NocT) and points toward the opening of the cleft by making six hydrogen bonds with the conserved side chains of residues Glu30/36 and Gln159/165 and the carbonyl of Ala88/94. Its carboxyl moiety makes a salt-bridge with the conserved Arg96/102 and three hydrogen bonds with the Ser91 side chain (the corresponding residue in NocT is Gly97) and the amide NH protons of Ser91/Gly97 and Thr163/Ser169. Its amide NH proton interacts with the carbonyl of Ala89/95 and the side chain of Arg96/102. The difference on the arginine moiety binding between Ocj and

Table 3
Comparison of Ocj and NocT structures and differences between the two structures with the ligand through hydrogen bonding interactions...

Table 3

Comparison of Ocj and NocT structures and differences between the two structures with the ligand through hydrogen bonding interactions...

Protein Residues

Ser91/Gly97	—	P35120	UniProt
Gly97	—	P35120	UniProt

Annotation source: West-Life Protein Residue Annotator

Screenshot of SciLite Annotations on europePMC; Firth et al. 2019

Show annotations in this article

Protein Residues (75)

- Ser91 (1/8) ...
- Asp161 (6) ...
- Gly97 (4) ...
- Asn202 (4/4) ...
- Gln122 (3) ...
- Ala164 (3) ...
- serine at position 97 (2) ...
- Ser92 (2) ...
- Glu30, (1) ...
- Gly97, (1) ...
- Ser202 (1) ...
- Asn202, (1) ...





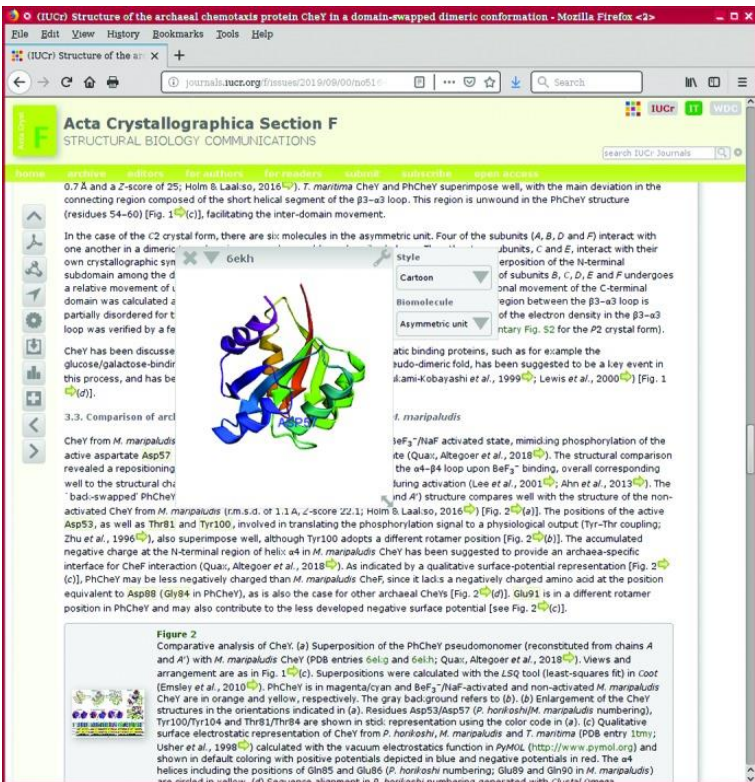
Europe PMC

Example of **excellent** practice for TDM developers and end users

- Submission via Web-Portal or API
 - Validation tool available
- Scope to host tools on EMBASSY Cloud
 - possible to run daily annotation jobs
- Annotations themselves are available under API
- Option of Bulk download
- Fulltext searchable and well formatted when retrieved as JSON
- Big infrastructure – ELIXIR Node
 - Not necessarily unique

Revamped platform imminent!





(IUCr) Structure of the archaeal chemotaxis protein CheY in a domain-swapped dimeric conformation - Mozilla Firefox <2>

Acta Crystallographica Section F
STRUCTURAL BIOLOGY COMMUNICATIONS

0.7 Å and a Z-score of 25; Holm & Laali-so, 2016^[10]). *T. maritima* CheY and PhCheY superimpose well, with the main deviation in the connecting region composed of the short helical segment of the β3-α3 loop. This region is unwound in the PhCheY structure (residues 54-60) (Fig. 1^[10](c)), facilitating the inter-domain movement.

In the case of the C2 crystal form, there are six molecules in the asymmetric unit. Four of the subunits (A, B, D and F) interact with one another in a dimeric crystallographic subdomain among the dimeric subunits. A relative movement of the domain was calculated as partially disordered for the loop was verified by a Fourier map. CheY has been discussed in the context of glucose/galactose-binding proteins, such as for example the *Staphylococcus aureus* glucose/galactose-binding protein (Lee et al., 2001^[11]; Ahn et al., 2013^[12]). The *Staphylococcus aureus* glucose/galactose-binding protein structure, which has been suggested to be a key event in the activation of the protein, has been suggested to be a key event in the activation of the protein (Lewis et al., 1999^[13]; Lewis et al., 2000^[14]) [Fig. 1^[10](d)].

3.3. Comparison of archaeal CheY from *M. maripaludis* and *P. horikoshii*. The active aspartate Asp57 revealed a repositioning well to the structural change in the 'bad-swapped' PhCheY structure. The archaeal CheY from *M. maripaludis* (r.m.s.d. of 1.1 Å, z-score 22.1; Holm & Laali-so, 2016^[10]) [Fig. 2^[10](a)]. The positions of the active Asp53, as well as Thr81 and Tyr100, involved in translating the phosphorylation signal to a physiological output (Tyr-Thr coupling; Zhu et al., 1996^[15]), also superimpose well, although Tyr100 adopts a different rotamer position [Fig. 2^[10](b)]. The accumulated negative charge at the N-terminal region of helix α4 in *M. maripaludis* CheY has been suggested to provide an archaea-specific interface for CheF interaction (Quax, Altegoer et al., 2018^[9]). As indicated by a qualitative surface-potential representation (Fig. 2^[10](c)), PhCheY may be less negatively charged than *M. maripaludis* CheY, since it lacks a negatively charged amino acid at the position equivalent to Asp88 (Gln84 in PhCheY), as is also the case for other archaeal CheYs [Fig. 2^[10](d)]. Gln91 is in a different rotamer position in PhCheY and may also contribute to the less developed negative surface potential [see Fig. 2^[10](c)].

Figure 2
Comparative analysis of CheY. (a) Superposition of the PhCheY pseudomer (reconstructed from chains A and A') with *M. maripaludis* CheY (PDB entries 6elg and 6elh; Quax, Altegoer et al., 2018^[9]). Views and orientations are as in Fig. 1^[10](c). Superpositions were calculated with the LSQ tool (least-squares fit) in Coot (Emsley et al., 2010^[16]). PhCheY is in magenta/cyan and BeF₃⁻/NaF-activated and non-activated *M. maripaludis* CheY are in orange and yellow, respectively. The gray bad-ground refers to (b). (b) Enlargement of the CheY structures in the orientations indicated in (a). Residues Asp53/Asp57 (*P. horikoshii*/*M. maripaludis* numbering), Tyr100/Tyr104 and Thr81/Thr84 are shown in stick representation using the color code in (a). (c) Qualitative surface electrostatic representation of CheY from *P. horikoshii*, *M. maripaludis* and *T. maritima* (PDB entry 1lmy; Usher et al., 1998^[17]) calculated with the vacuum electrostatics function in PyMOL (<http://www.pymol.org>) and shown in default coloring with positive potentials depicted in blue and negative potentials in red. The α4 helices including the positions of Gln85 and Glu86 (*P. horikoshii* numbering); Glu89 and Gln90 in *M. maripaludis*) are depicted in yellow. (d) Sequence alignment of *M. maripaludis* CheY with *P. horikoshii* numbering overlaid with *T. maritima* CheY.

Pyresid now runs on ingestion to *Acta Cryst. F*

Text Source independent of ePMC
Implemented by IUCr Developer
Simon Westrip

Existing Acta Cryst F. archive has been annotated

Linked with 3Dmol.js to annotate 3D Structure

Figure: van Raaij & Newman; Taking biological structure communications into the third dimension. *Acta Crystallogr F Struct Biol Commun.*



Structure of the archaeal chemotaxis protein CheY in a domain-swapped dimeric conformation

Karthik Shivaji Paithankar,^{a†} Mathias Enderle,^{a†} David C. Wirthensohn,^{b†} Arthur Miller,^b Matthias Schlesner,^b Friedhelm Pfeiffer,^c Alexander Rittner,^a Martin Grininger^{a*} and Dieter Oesterhelt^{b*}

^aInstitute of Organic Chemistry and Chemical Biology, Buchmann Institute for Molecular Life Sciences, Goethe University Frankfurt, Max-von-Laue-Strasse 15, 60438 Frankfurt am Main, Germany, ^bDepartment of Membrane Biochemistry, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany, and ^cComputational Biology Group, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

*Correspondence e-mail: grininger@chemie.uni-frankfurt.de, oesterhe@biochem.mpg.de

Edited by N. Sträter, University of Leipzig, Germany (Received 23 April 2019; accepted 4 August 2019; online 30 August 2019)

Archaea are motile by the rotation of the archaellum. The archaellum switches between clockwise and counterclockwise rotation, and movement along a chemical gradient is possible by modulation of the switching frequency. This modulation involves the response regulator CheY and the archaellum adaptor protein CheF. In this study, two new crystal forms and protein structures of CheY are reported. In both crystal forms, CheY is arranged in a domain-swapped conformation. CheF, the protein bridging the chemotaxis signal transduction system and the motility apparatus, was recombinantly expressed, purified and subjected to X-ray data collection.

Keywords: chemotaxis; signal transduction; response regulator; CheY; CheF; archaellum; protein evolution.

PDB references: PhCheY, 6er7; 6exr

[Similar articles](#) [PowerPoint slides](#)

1. Introduction

Archaea and bacteria share the ability to move in response to chemical or physical stimuli towards favorable growth conditions (Marwan & Oesterhelt, 2000; Quax, Albers *et al.*, 2018). Motility is based on the rotation of the flagellum (in bacteria) and the archaellum (in archaea; formerly known as the archaeal flagellum), respectively, and the directionality of the movement is provided by modulating the switching frequency in response to the stimulus (Armitage, 1999). The molecular basis underlying taxis is composed of two systems: chemotaxis signal transduction, which processes the external stimulus, and the flagellum/archaellum, which responds to the chemotaxis output signal.

The Che proteins, encoded by genes that cluster in genomes, constitute the chemotaxis signal transduction system. The overall mechanism of chemotaxis is conserved in archaea and bacteria (Szurmant & Ordal, 2004). Receptors, generally known as methyl-accepting chemotaxis proteins (MCPs) and referred to as halobacterial transducer proteins (Htrs) in halophilic archaea (Zhang *et al.*, 1996), sense external stimuli such as chemicals, oxygen or light. The histidine kinase CheA and the response regulator CheY form a stimulus-response coupling mechanism, generally termed the two-component system (Parkinson & Kofoid, 1992; Parkinson, 1993). CheA autophosphorylates and subsequently donates the phosphate to CheY, yielding phosphorylated CheY (CheY-P; Garrity & Ordal, 1997; Bischoff *et al.*, 1993; Rudolph & Oesterhelt, 1995; Rudolph *et al.*, 1995). The concentration of CheY-P determines the switching frequency of the flagellum or archaellum, respectively. Several Che proteins are involved in adapting (CheR, CheB, CheC, CheD and CheV; Springer & Koshland, 1977; Simms *et al.*, 1985,

Thoughts

- What is made available elsewhere? Will your corpus be disregarded?
 - Don't reinvent the metadata wheel e.g. W3C standards: Web Annotation Data Model etc.
- Table Schemas in Metadata
 - Semi-structured data is good target, but scaled extraction can be challenging without enforcing standards
- Closing the loop
 - Make available other mining outputs
 - *Version* these mining outputs – is it up to date?
- Capture Preferred Vocabularies and Ontologies
- Corrections can be powerful
- Validation and Evaluation
 - Valid format?
 - Valid annotations?



Enabling Annotation

- A well annotated corpus lends itself to further exploitation
- More “Classic” TDM: Named Entity Recognition, Information Extraction
- Other, broader tasks like Question Answering and Summarisation
- Latest technology in Natural Language Processing, ‘Transformers’ (Neural Net Architecture)
 - “Imagenet for Text” – BERT/GPT-2 pretrained models
 - Transfer learning in this area is already being done: “*BioBert: a pre-trained biomedical language representation model for biomedical text mining*” Lee et al. 2019



Evaluation

- Peer review gives unparalleled access to expert knowledge
- Most expensive part of building ML/AI tools is often expert Humans, not CPU/GPU hours
- Authors have **expectation** of *time* spent during submission
 - This is unusual!
- Evaluation of Text Mined Terms and relationships
 - Put back the results to submitters
 - Refine the model, provide a better offering
- More in the next session!





Hartree Centre

Science & Technology Facilities Council

Thanks!

Find out more:

robert.firth@stfc.ac.uk

@ hartree@stfc.ac.uk

 hartree.stfc.ac.uk

 [/company/stfc-hartree-centre](https://www.linkedin.com/company/stfc-hartree-centre)

 [@hartreecentre](https://twitter.com/hartreecentre)



Hartree Centre

Science & Technology Facilities Council

Additional Slides



Hartree Centre
Science & Technology Facilities Council

Citations

Firth R, Talo F, Venkatesan F, Mukhopadhyay A, McEntyre J, Velankar S, Morris C, *Acta Crystallogr F Struct Biol Commun.* 2019 Nov 1; 75(Pt 11): 665–672. Published online 2019 Nov 5. doi: 10.1107/S2053230X1901210X, PMID: PMC6839820

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, 09-2019, btz682, <https://doi.org/10.1093/bioinformatics/btz682>

van Raaij MJ, Newman J. Taking biological structure communications into the third dimension. *Acta Crystallogr F Struct Biol Commun.* 2019;75(Pt 11):663–664. doi:10.1107/S2053230X19014754

Venkatesan A, Kim JH, Talo F, et al. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data. *Wellcome Open Res.* 2017;1:25. Published 2017 Jul 10. doi:10.12688/wellcomeopenres.10210.2

Rego, N. & Koes, D. (2015). *Bioinformatics*, **31**, 1322–1324.



Additional Examples - AstroCats

The Open Supernova Catalog

Catalog About Contribute Derivations Statistics Download Bibliography Links


BVRI LCs added for SN2012ap <https://t.co/Mpaov618ZP> 6 hours ago

Astro Catalogs @AstroCatalogs

We obtained an optical spectrum [range 370-845 nm] of SN 2016fbz (= Gaia16bbj) on UT Aug.27.6.2016 with the 2.16-m... <https://t.co/Cw8WmT5j> 11 hours ago








































ATel @astronomeratel

Welcome to the open supernova catalog! The goal of this catalog is to act as a centralized, open repository for supernova metadata, light curves, and spectra. The data on this page is scraped from various supernova data repositories, both defunct and active, and from individual papers that have published their data in machine-readable form. If you use this data, please reference the cited sources of that data. We'd also appreciate if you referenced the paper describing this catalog. Thanks!

The table below is auto-updated from a [GitHub repository](#) which encodes the data on each event as a series of ASCII files in [JSON format](#). The entirety of the data available for any supernova can be downloaded by clicking the  icon in the Data column. If you would like to contribute data yourself, please visit our [contribute](#) page. If you are aware of a source of data that is already available either online or in the literature, please add the source of data to our [to do list](#). If you spot any mistakes, please create a new issue on our [GitHub issue tracking page](#), or contact us via e-mail.

Select all Deselect all Column visibility Export selected to CSV Search:

Show 10 entries Previous 1 2 3 4 5 ... 3651 Next

Name	Disc. Date	m_{max}	Host Name	R.A.	Dec.	z	Type	Phot.	Spec.	Radio	Data
<input type="checkbox"/> SN1987A	1987/02/24	4.53	 LMC	05:35:28.020	-69:16:11.07	9.51e-06	II Pec	 3332	 36		
<input type="checkbox"/> SN2011fe	2011/08/24	9.893	 NGC 5457	14:03:05.711	+54:16:25.22	0.000804	Ia	 2735	 85	 0	
<input type="checkbox"/> SN2003dh	2003/03/31	14.64	 A104450+2131	10:44:50.01	+21:31:17.8	0.1685	Ic BL	 2687	 13		
<input type="checkbox"/> SN1993J	1993/03/28	10.77	 NGC 3031	09:55:24.7747	+69:01:13.702	-0.000113	I Ib	 1815	 50		
<input type="checkbox"/> SN2002ap	2002/01/29	12.72	 NGC 628	01:36:23.85	+15:45:13.2	0.002108	Ic BL	 1781	 39		
<input type="checkbox"/> SN2009ip	2009/08/26	13.73	 NGC 7259	22:23:08.26	-28:56:52.4	0.005944	I In	 1569	 237		
<input type="checkbox"/> SN2000cx	2000/07/17	13.39	 NGC 524	01:24:46.19	+09:30:31.3	0.007929	Ia Pec	 1300	 45		
<input type="checkbox"/> SN1999em	1999/10/29	13.68	 NGC 1637	04:41:27.04	-02:51:45.2	0.00223	II P	 1172	 70		
<input type="checkbox"/> SN2011dh	2011/06/01	13.32	 NGC 5194	13:30:05.1055	+47:10:10.922	0.001638	I Ib	 1122	 78		
<input type="checkbox"/> SN1999ee	1999/10/07	14.93	 IC 5179	22:16:09.40	-36:50:31.5	0.01141	Ia	 1102	 26		

Name Disc. Date m_{max} Host Name R.A. Dec. z Type Phot. Spe. Rad

Showing 1 to 10 of 36,503 entries Previous 1 2 3 4 5 ... 3651 Next

Last modified August 29 2016, 02:08:34 [UTC].

Of the 55,430 supernovae that have been discovered, only:

- 18,172 have publicly available light curves in an easily downloadable format
- 8,600 have spectra



