Protein structure prediction using homology modeling [Edit]

From Learn Bioinformatics course • 13 min read

What are proteins?

Proteins are large biomolecules which are responsible for performing most of the functions within an organisms cells, including responding to stimuli, acting as catalysts for other reactions, transporting molecules from one place to another and performing cell signaling. Just like DNA sequences, protein sequences are *strings of molecules* but unlike DNA sequences, there are 20 different molecules called *amino-acids* that make up protein sequences.

Protein structure

Every 1D protein sequence string folds into **3D structures**. These 3D protein structures are determine how a protein responds to various environments and which other molecules it interacts with, and hence is critical in the ability of the protein to perform its functions. The 3D structure of protein is described by providing the *coordinates* (*x*-*y*-*z*) of every *atom* in the protein, in *3D space*.

Determining protein structure

Protein structures can be determined using experimental procedures like X-ray crystallography and Nuclear Magnetic Resonance (NMR). However, these techniques are slow and cumbersome, and cannot be applied to all the proteins. Therefore, *high-throughput* computational methods are used to *predict* 3D structures of proteins from sequences.

Homology Modeling

One of most popular computational methods for protein structure prediction is *Homology Modeling*. Homology modeling leverages the property of *evolutionary conservation* of https://www.commonlounge.com/discussion/73dd4037cc00473fb37bf4c52ec5bf30#_=_

. . .

homology modeling, this property of conservation of protein structure is used to predict structures of newly discovered protein sequences whose structures cannot be resolved using traditional experimental methods.

The main idea is that protein sequence with unknown structure is searched against the sequence database of proteins where structures of all proteins are experimentally known and the unknown structure is *modeled* from the *evolutionarily* closest or best matching protein from the database.

In this article, we describe the approach and methodology of homology modeling, i.e. how it works. We also describe how to use the SWISS-MODEL tool for performing homology modeling.

Detailed description of Homology Modeling method

In this section, we will provide an overview of the steps involved in homology modeling. Note that a number of these steps are active areas of research.

An mentioned previously, homology modeling starts with knowledge of the structure of a number of proteins and their sequences which have been determined by experimental methods. The method uses this previous knowledge to predict the structure of proteins for which we know the sequence, but don't yet know the 3D structure.

For predicting the structure of the protein, we'll first predict the co-ordinates of N, C_a , C_b (backbone) and then the co-ordinates of the R-group (side-chain) of each amino acid.



Step 1: Template recognition and initial alignment

First we find the evolutionarily closest proteins to the target (protein for which we want the predict the structure). This is achieved using database search algorithms like BLAST (Basic Local Alignment Search Tool) which perform sequence alignment of the target sequence against the database of proteins sequences. The PDB (Protein Data Bank) is one such database. The best matching protein sequence from the database, to our target is assumed to be the evolutionarily closest and its structure will be used as a **template** to the model the structure of the target. The database search tool also gives an alignment, i.e. information of which regions of the target match which regions of the template.

Step 2: Alignment correction

The initial alignment between the target and template obtained during the database search may not be optimum in certain difficult regions of the alignments. For example, the initial alignment may violate some rules of amino-acid substitutions like substituting hydrophilic residue with hydrophobic residue in the core of the protein.

Given that we have already found an initial template, we can now use more expensive alignment algorithms to find a better alignment. For example, we can use multiple sequence alignment algorithms for this step. Multiple sequence alignment are useful for identifying regions that are highly divergent, and hence better detecting the appropriate locations for insertions and deletions.

Step 3: Backbone generation

After the target-template alignment is optimized, the protein backbone structure (N- C_a - C_b) for the target is generated. This is achieved by simply copying the coordinates of the template backbone to the target based on the alignment. That is to say, the coordinates of an atom in the target protein is said to be the same as the coordinates of the corresponding atom in the template protein, as suggested by the alignment from the previous step. Obviously, this process is highly dependent on the accuracy of the template structure, and any errors in our initial database will cause errors in our prediction.

alignment, insertions and deletions. Incorporating these mismatches into the backbone is the most difficult part in homology modeling.

The secondary structure of the protein consists of helices, strands and loops. Since conformational changes implied by insertions and deletions can't happen in helices and strands, they must happen in the loops.

There are two main approaches to model loops: knowledge based and energy based. The former approach searches for conformations of loops that have similar sequences and endpoints as the target in the database of known structures. The latter models the loop conformation in an ab initio manner by predicting the loop structures with lowest structural energies using force field functions and molecular dynamics. These methods provide fairly accurate results for short loops with up to 5-8 residues.

Step 5: Side-chain modeling

Modeling the side chains involves predicting the value of C_a - C_b torsion angle for each Rgroup attached to the backbone. The conformation of side-chains in the structures, also called the rotamers, depends on the values on this torsion angle. The side-chain are generally modeled in a knowledge based manner using rotamer libraries which contain preferred conformations for all 20 R-groups under various chemical neighborhoods.

Step 6: Model optimization

Now that all the aspects of protein structure are modeled for the target, it is time make fine changes in the structure to reduce the overall energy. This is achieved in an iterative manner. In each iteration, the backbone conformations and rotamer conformations are changed alternatively to lower the overall energy of the predicted structures.

Model optimization can also be performed by running a molecular dynamics simulation, which starts with the current predicted structure, and makes small changes to the structure based on the simulation, i.e. we simulate what would happen to each of the atoms of the protein under the forces surrounding it on a femtosecond (10⁻¹⁵) timescale.

me mai step is to check the predicted structure for errors. Errors are introduced in the predicted protein structures due to low alignment between target and template or due to errors in template structures. Checks are performed on the predicted structure to see if all the bond lengths, bond angles and torsion angles fall in characteristic ranges found from experimentally determined protein structures. Energy checks are also performed which see if the different types of structure based energies like Van der Waals and Electrostatic are at expected levels.

Homology modeling using SWISS-MODEL

We will take detailed look at homology modeling procedure by predicting structure for protein *Ornithine carbamoyltransferase* (UniProtKB accession: P96134) present in bacteria *Thermus thermophilus* using the SWISS-MODEL $rac{1}$ tool.

Target-template recognition

First step is to search the target sequence against the database of sequences with known protein structures. Paste accession number into the window and hit the "Search For Templates" button

rget Sequence(s)	Faste your target sequence(s) or ThiFrotRB AC here	Supported Inputs @
ormat must be 74514, Ciustel, in string, or a valid UniProtVB		Sequence(s) +
9		Tarpel-Template Alignment +
		User Template +
	A United Toront Descent a Line Relationship	DeepView Project +
	The second residence and the second s	
yed the	Untitled Project	
nak.	Optional	
	Search For Templates Build Model	
	By using the StirtSS-MOCEL server, you agree to comply with the following terms of use and to oils the corresponding articles.	
tart a New Modelling	By using the SRISS ACCEL sener, you agree to comply with the following terms of use and to ofe the corresponding addres.	Constant look of
tart a New Modelling arget Sequence(s): treat-suit ar https://cuaix/		Supported Inputs @
Sart a New Modelling aget Sequence(s): umatinual be PKI54, Chank, in atting or a safe lisihoot8 0	Project Pr	Supported Inputs @ tequence(s) *
Rart a New Modelling aget Sequence(s); manimus te reformand an aning, or a valor lasheed O	Project e Targen Residence Project e Targen Residence Residence Residence Residence Residence Residence Residence Residence Residence Residence Residence Residence Residence Residence Residence Residence Residence	Supported Inputs @ Tequences) * Target Templas Algoment *
Bart a New Modelling arget Sequence(s); none water entity, chang, an uting, or a valer lawhard O	By using the SONSS ACCEL, server, you agree to comply with the billiouning terms of use and to oils the corresponding adcess Project Project Project Intervent Intervent	Supported Inputs @ Bequence(t) = Target Temptals Alignment = User Temptals =
Start a New Modelling arget Sequencess: man mate in ACM, Classe an aring, or a safe (softward) CO		Supported Inputs Tequencess Teque Template Algoment = User Template = Despites Poped =
Rant a New Modelling arget Sequence(s): formar wait ar MDM, Classe, an arms, or a valor looked O	By using the BRITSS ACCEL, server; you agree its comply with the bilinuing terms of use and to oils the corresponding access Project Project Concept Destination of the server of the bilinuing terms of use and to oils the corresponding access Project Destination Destin Destination Destination Destination Destination	Supported Inputs Tequencess Tequet Sergistic Algorism User Template Complete Poped =
Start a New Modelling angel Sequence(s): unit arting or a salar biofload o noject Title: mail:	By using the BottSEACCEL server, you agree to comply with the billowing terms of use and to obt the corresponding address Project © Torgets Content Tore Hill a superior to content and a content to content t	Supported Inputs Sequence(i) = Target Temptate Adjorment = User Temptate = Desplifeer Project =

how well their sequences align with the target protein sequence. The first selected structure template is the best matching (99% identity). The second template matches the target sequence with 51% identity. The superimposed protein structures of the two top matching templates can be seen in the window on the right.

We will use the top two results to build or predict two structures for the target sequence and then select the best predicted structure.

Tem	plate I	Results e														
Tem	plates	Quaternary Structure Sequence Similarity	Algoment of Selected	Templater	More	-					Bi	Build Model	Build Models 🕑	Build Models (3)	Build Models 📀	Build Models 🕘
	• Name	• 16e	Coverage	+ Identity	• Method	Oligo State	Ligands					Clear Select	Clear Selection	Clear Selection	Clear Selection	Clear Selection
	2e6.1.A	Omithine carbamoytransferase		99.34	X-ray 2.0Å	monomer 🗸	1×80.0	*								
	3935.1.4	Omithine-carbamoytransferase		51.19	X-18(2.1Å	homo-hexamer /	None	~								
	1pw1.4	Omithine carbamoythansferase		47.48	X-ray, 1.9Å	homo-12-mer J	None	¥								
	1pw1.A	Cmithine carbamoy/transferase		49.48	X-18(1.9Å	homo-12-mer J	None	¥			<u></u>					
	1416.1.4	ORNITHINE CARBAMOVLTRANSFERASE		49.48	X-ray 2.7Å	homo-12-mer /	None	٣			621					
	fats.1.4	ORNITHINE CARBAMOVLTRANSFERASE		47.96	X-ray 2.7Å	homo-12-mer J	None	¥				C 22				
	4601.1.4	Omithine carbamoy/banoferase		44.90	X-ray 3.2Å	homo-bimer 2	None	¥			20	2020				
	4017.1.4	Omithine carbamoy/bansferase		44.25	X-ray 1.5Å	hemo-kimer /	None	٣			2					
	412.1.4	Omithine carbamoy/bansferase		44.55	X-10, 17Å	homo-birmer 4	3 x OP ¹⁰ , 3 x N/A ¹⁰	٣								
	4857.1.4	Omithine carbomoylitansferase		45.05	X-18, 1.5Å	hemo-bimer 2	None	٣								
	dep1.1.4	Omithine carbamoyttansferase		45.21	X-ray, 3.2Å	hamo-bimer 2	None	٣			. 📕					
	156.1.A	ORNITHINE TRANSCAREARMLASE		41.36	X-18(2.6Å	hemo-timer /	3 x OP 7	٣			· · · ·				· · · ·	· • • • • • • • • • • • • • • • • • • •
	1dly1.A	ORNITHINE CARBAMOVLTRANSPERASE		41.36	X-ray, 1.0A	here-biner 2	3 × CP ⁻⁷ , 3 × N/A ⁻⁷	¥								
	185.1.4	ORNITHINE TRANSCAREABOYLADE		42.18	X-ray, 3.5Å	homo-bimer /	3 x NVA rd	٣				1-	1-		1-	1-
	150.1.A	ORNITHINE TRANSCAREAIMLASE		43.45	х-на; 2.6Å	hemo-timer /	3 x OP 7	¥								
	1cBy1.A	ORNITHINE CARBAMOVLTRANSPERASE		43.45	X-18, 1.9Å	homo-birner J	3 + CP ⁻¹ , 3 + N/A ⁻¹	¥	240	240.14	24014	240.14	24014	2+014 (2+01A	2001A
	14(1.4	Cmithine carbamoythansferase		40.82	X-18(2.2Å	homo-12-mer J	None	¥	390	3p05.1.A	3gd5.1A	3pd5.1.A	3pt51A	3p85.1A	3pt5.1.A	39051A C

The modeling results can be seen below. The two predicted structures are ranked according to the quality of their models. Next, we will assess the quality of both the predicted structures to see which is the best.

S	Golde Costilly Extension 0.37 C/D -0.55 All Atom -0.25 Schwalton 0.59 Torsion 0.40 Template Seg Meetity 2vit0.1A 99.34% Hoded Template Alignment	Lee d' Guelly L'estevalue		Citck model image to view in 3D
	Citigan-State @ Ligands @ Homo Home hexamer (matching prediction) Gother Desiting Enterance QMEAN - 1.20 Cp - 1.20 Cp - 0.00 Al Atom - 0.25 Solvation - 1.107 Template Seq.Meeting Coverage 3935 1 A 51.10% Hoded-Template Alignment	Local Quality Estimate	Comparison	Try & Cattor & A > A O

The QMEAN is one of the primary measures used to assess the model quality. QMEAN is a composite scoring function based on different geometrical properties of protein structures and provide both global (*i.e.* for the entire structure) and local (*i.e.* per residue) absolute quality estimates.

QMEAN consists of four individual terms. The four individual terms of the global QMEAN quality scores are also listed. The white area in the bar-plots (numerical values close to zero) indicates that the property is similar to what is observed in experimental structures. Positive values indicate that the model scores higher than experimental structures on average, negative numbers indicate that the model scores lower than experimental structures on average.

For the first model (built using 2ef0.1.A as a template) the QMEAN terms mostly fall within the white region.

33	CMEAN 0.37 C0 0.37 Al Alom 0.55 Solvation 0.59 Torsion 0.59	1 nontriget, solar lane,				
	Template Seq.Identity Coverage 2000.1.A 90.34%	Description Omithine carbamoytransler	254	_		
	Biounit Oligo State Method monomer X-ray, 2	Seq Similarity 2.00 Å 0.59	Range Coverage 7 - 301 1.00			
	Ligand Added to Model MG X - Not biologically releva NA X - Not biologically releva	ant.	Description MAGNESIUM ION SODIUM ION		Ser and a series of the series	
	Model Template Alignment			•	PY . Catron	
Nodel_CLEOSTAL	TOPATLOALLOLAT LANDA TROEDLAND VIA	LLPERFELETETTLEVANVELOGRAVE	DOROWGEOREE PVRDVARE LERF	1		301
Nodel_CL WEGTARRWERKETVERLE 2010.1.A.VIGTARRWERKETVERLE	ARATVPVNALSDRANPLQALADILTLENVEN AREAVVENVERŽEDRASČI LQALADILITLEVEN	OCLACIEVANVORCHEVLES LLEVAPL	CLEVENT PROTECTION			
Nodel_11 APPTHOPPERALCANAL	TRANSFERROR AND AND ADDRESS OF TAXABLE PROPERTY AND ADDRESS OF TAXABLE PROPERTY AND ADDRESS OF TAXABLE PROPERTY AD	LIRPEONPLECIPARTORETTERAVEG	PREBVPDGARBELETAKAVLATLE			
server and the state of the server of the se			Property I to gath I it and a shift at a large			
Nodel_ti 2efi.l.k/K				665. 1		
Rest, S. B RET. J. A.X	Oligo-State @ Monomer	Ligands @ None		80	GMQE 0.97	0 QMEAN 00 0.37 №7
	Oligo-State Monomer Odel 01	Ligands () None Estimate () 0.37 (-0.55 () 0.20 () 0.59 () 0.40	Local Quality Er	itimate	GMQE 0.97	QMEAN 0 0.37 K ⁵
	Oligo-State Monomer Global Quality E QMEAN Cβ All Atom Solvation Torsion Template Set0 1 A 9	Ligands None Estimate 0.37 0.55 0.59 0.40 ieq Identity Coverage 19.34%	Local Quality Er	timate	GMQE 0.97	CAMEAN O 0.37 KS

However, for the second model (built using 3gd5.1.A as a template) most QMEAN terms significantly differ from the optimal

	Otgo-State Ø Lagends Ø Homo- None hexamer (matching prediction)		GMQE Ø GME 0.75 -1.20	AH Đ SIÔ	
	Gateban Color -1.20 -0.60 -0.60 -0.60 -0.25 -0.25 Solvation -0.25 -0.15 -0.15 -0.15 -0.167			^	
	Template Seq Identity Coverage Ogd5.1.A 51.19%	Description Ormithine carbamoyltransferase		^	
	Bisenit Olgo State Metho homo-hesamer X-ray,	d Seq Similarity 2.10 Å 0.43	Range Coverage 6 - 301 0.98		
	Nodel-Template Alignment			^	
OModel_02:AMOOTASTEPROLLOFSO	OPPE LOALLD LASPLEY BRYDOS PLEORY	LALLFREPS LETETTLEVANVELOGEAVTL	OROVOTOTRA PURCHASING AN	1 1 1 1	
Nodel_0218 NGGEALTERNOLLDYSC	SPEELOALLPLAERLERSPIECEPLECEV	ALLEREPS CREATE THE VANYELOG BAYTLE	ORDVOICEREPVRDVARELER	100	
Hodel_021CHOGEALTERBOLDFSG	OPER-LOAL LO LARBLED RETRORDLECTV	LALLEREPS LETETTLEVANVELOGEAVTL	GEOVO LORREPVEDVARE LER	100	- Column - A - A - A - A
Noted ALL MORE THE PERSON AND AND A	CPER LOAD LABOR LEVEL STREET SOLE CENT	LALLY TO PERFORM TO PARTY LOOP AVT L	COMPANY OF THE PARTY NOT A COLUMN		ALL CHORE A P A D
Nodel GITTMSGEALTERN LDFSG	OPPE LOAL LO LARRIERE TREET LE CENTRE	LALLFEETSLETTTLEVANVELOGEAVTL	GEOVOIGER PVROVAKELER	1.00	1
5pt5.1.2	OTAGLEALLTLANGLERO-TRVATLOORS	LOLVELES STREETS OF TWANT CLOSED VIOL	POSTO OF REPARTANCE IS	1 100	
Hode) 03:2 VEGEARVERNETVEAL	BRANNPOWERLSBRAMPLOALADUSTUM	MPOSTAGLEVANVGDGBEVLEST EVAPLA	LEVEVATPROTEPOPOLLER.	- 100	
Hodel 0218 VEGEAARVERSTVEALS	REARVPVVRALSDRARPLGALADLETERS	VPGGLAGLEVANVCDCREVLES LLEVAPLAS	LEVEVATPROTEPPPOLLER	1.00	
Hode1_02+C VEGEAARVEREETVEAL	REARYPVVBALSDBARPLOALADLUTURY	VPOCLACLEVANVODCESVLESLLEVAPLA	LEVEVATPROTEPOPOLLER	- 100	
Hode1_03+0 VEGEAARVERSETVEAS	BRARVFVVNALSDBARFUGALADUUTURS	VF GOLAG LEVANYOD GERVLES LLEVAPLAS	LEVENA TPROTEPOPOLLER	- 180	
Model_021E VEGEAARVEREETVEAL	BRARVPVVRALSPRARFUGALADLUTURS	VF66LAGLEVANVED-DEFVLESLLEVAPLAS	LEVENATPEGTEPOPOLLER	- 188	
Nodel_0217 VEGEAARVEREETVEALS	BRARVPVVNALSDRARPUGALADUUTURS	IVF GELAS LEVANVOD GREVLES E LEVAP LAS	LEVENAT PROTEPOPOLLER-	- 189	

0801100						
	Sec. Sec.	MODELUZ .	prediction)			
			Global Quality Estimate QMEAN -1.20 Cβ -0.60 All Atom -0.25 Solvation 0.15 Torsion -1.07	Local Quality Estimate	Comparison	^
			Template Seq Identity Coverage 3gd5.1.A 51.19%	Description Ornithine carbamoyltransfera:	se	~
			Model-Template Alignment			۲

Therefore, the structure predicted by the template 2ef0.1.A is the most optimal model and can be used as the *predicted structure* for our target sequence.

Model Results e			Order by: GMQE	v	
14 A	Oligo-State Ligands Monomer None		GMQE 0.97	QMEAN 0.37 KS	(SA
	Global Quelly E stimate GMEAN Cβ All Aban Clip Clip	Local Gestily Estimate		Î	
	Template Seq identity Coverage 2etD 1.A 99.34%	Description Omithine carbamoytransferase	Berlind	v	
	Model-Template Alignment			^	
2#f0.1.AHOOKALTLPHDLLDFSOYO	PWELGALLOLAEGLKRERYBOEDLKOKVLALLPE	EPSLATATTLEVANVELOGEAVYLD;	NOVOIOEREPVRDVAKSLI	EB.F 100	
Nodel_01 VEGEAARVFRAETVEALAR 2#20.1.8 VEGEAARVFRAETVEALAR	HARVPWRALKORARPLOALADLETLERVFORLA HARVPVYRALSORARPLOALADLETLERVFORLA	OLEVANVODORNVLESLLEVAPLAGI	NYRVAT PROTE POPOLINI NYRVAT PROTE POPOLINI	1.3.5° 200	
Model_01 AFFTHDFEEAALGARALIT 2mf0.1.AAFFTHDFEEAALGARALIT	NAME SHOOF A CRAINER OF DO FOR SHOELLKLER P	EGVFLECIPARTGEETTEEAVEGPRE EGVFLECIPARTGEETTEEAVEGPRE	SVYDDARMSLETAKAVLL SVYDDAENSLETAKAVLL	TLL and	-
Model_01				905. 905.	E PV a Cation a 🙆 🕨 📥 💭
					1 301

Resources

- MODELLER d'

References

• "Homology Modeling" by Elmar Krieger, Sander B. Nabuurs, and Gert Vriend

Protein structure Swiss-model Homology modeling Protein structure prediction Protein

Category: Biology • Last updated: 2 years ago



★★★★★ Leave a Review

< **PREVIOUS**

Detecting Mutations with Read Mapping and Suffix Trees

ABOUT THE CONTRIBUTORS



Akshay Yadav

Masters in Bioinformatics and 6 years of work experience in Bioinformatics & Computational Biology



Keshav Dhandhania

BSc, MSc 2014 @ MIT (AI, Deep Learning). Former Competitive Programmer.

Add a comment

Back to top

POPULAR PATHS & COURSES

Machine Learning Deep Learning Natural Language Processing Big Data Data Science Bioinformatics Algorithms

NEW PATHS & COURSES

Web Development UX & UI Design Startups Product Management Cryptocurrencies Finance College Admissions

ABOUT US

Our Mission
How to Contribute
Help and FAQ
CommonLounge Meta
Team
Privacy, Terms & Refunds

GET IN TOUCH

hello@commonlounge.com Facebook Twitter San Francisco, CA, USA +1 844 318 7406

Copyright 2016-2020, Compose Labs Inc. All rights reserved.