How to get functional information and biological insight from structural bioinformatics tools, predictors and public data – some suggestions

Getting the correct protein sequence

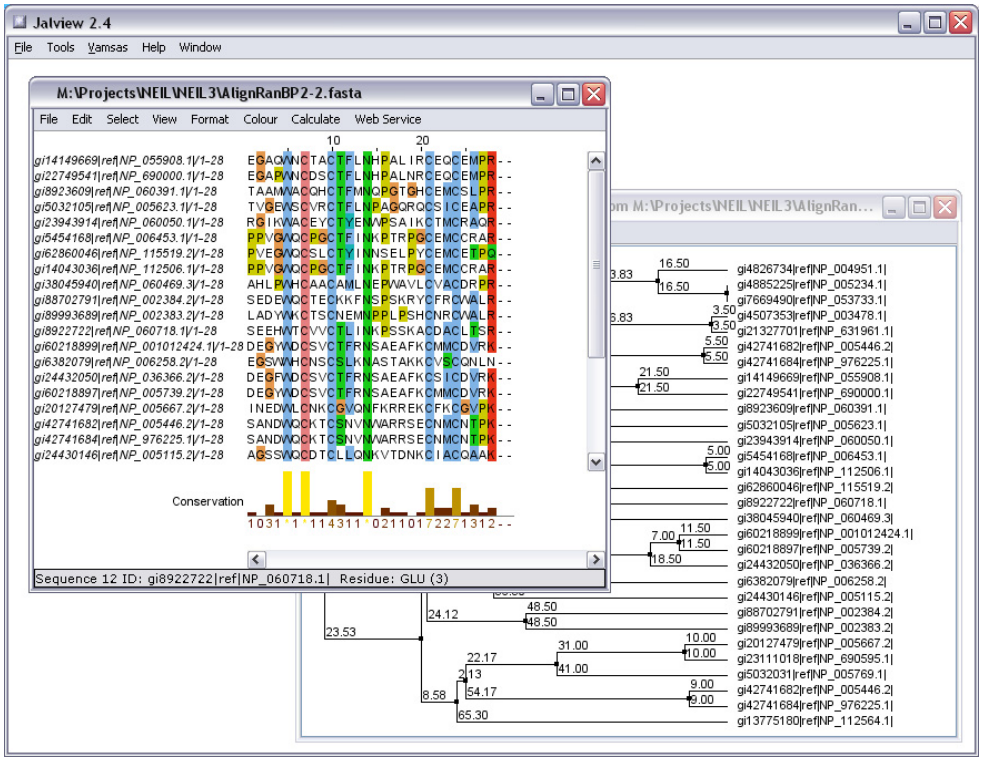
Surprisingly often it is not exactly clear what the protein sequence is. Before any modeling can be attempted it is absolutely necessary to get an idea about the correct sequence.

- Getting the correct sequence is usually easier in prokaryotes where most protein-coding genes have a single ORF. Still it may be unclear what the actual start codon is, and an ORF might in many cases not correspond to a functional protein.

- Extracting the correct sequence is often much more tricky in vertebrates, plants, and many other eukaryotes due to spliced genes and alternative splicing. mRNA, cDNA and EST sequences often do not correspond to transcripts that can be translated into a functional protein.

  - Transcriptional junk?
  - The sequences straight out of RefSeq, UniProtKB, or other databases are certainly not always correct.
  - It may be necessary to "fish out", from a huge amount of data, the most likely correct exons with correct splicing. Remember, if you are checking a human protein and looking for orthologs:

    - UniProtKB/Swiss-Prot database (http://www.uniprot.org): It is manually curated. A person (expert) has actually looked at the mRNA/protein and checked that it makes sense. Likely to find perhaps mouse, rat, zebrafish, *X. tropicalis* frog orthologs, but not many more sequences.
    - NCBI nr (http://www.ncbi.nlm.nih.gov/protein), RefSeq, UniProtKB/TrEMBL: The proteins are mainly automatically generated. Most likely no-one has ever looked at the data. Errors are usually not corrected. You are likely to find Swiss-Prot sequences and in addition bovine, opossum, chicken, fugu, and a few more vertebrates.
    - Ensembl (http://www.ensembl.org): The proteins are 100% automatically generated from often noisy genomic data. Lots of obvious errors such as single- or two-base introns that have been introduced by the software to get rid of frame-shift sequencing errors. Contains same species as UniProt/nr but not necessarily the same sequences. In addition orthologs from 35-40 mammals, frogs, chicken, turkey, and zebra finch, a lizard, a turtle, 9 fish species, sea lamprey, lancelet and sea squirts. Next year there will be perhaps 20 new species. In *Pre!*Ensembl you will find several more species, for example the recently releases flycatcher and the sheep. *It is this kind of data that is growing*!
    - Sequencing centers, sequencing projects websites, e.g. DOE JGI (http://genome.jgi-psf.org): Proteins also here mainly completely automatically generated. Even more non-chordate species like mollusks, insects, crustaceans, Trichoplax, sea anemone, sponges etc.
    - From this rich source of data it is usually possible to get a very good idea about what is the likely human sequence. ***You have to look carefully at the data and work with it!*** Exonic segments that are conserved in for example all mammals or all vertebrates are clearly under evolutionary selection and is "important". Splice variants or proposed exons found only in human/primate data but not conserved in any other species is much more likely to be "junk". An extremely useful tool for looking at and working with sequence data from 10s or 100s of homologs is Jalview (http://www.jalview.org).



Below is given an example of Ensembl data for the orthologs of human PCSK9. Many of the sequences lack the whole or parts of exon 1 and 2. At the bottom is shown the extracted, most likely "good" sequences from primates and mouse.

- Working with sequences like this may in many cases be more than 50% of the job!

- A multiple sequence alignment (MSA) of a human protein with for example 25 "good" orthologs from other vertebrates may in many cases be more useful (to get biological insight) than for example a 3D model. It might also make the model much more useful since it tells you which parts of the protein (residues, segments, or domains) are under selective pressure. Presence or not in various species may also give hints about function. You can read about a project were we extracted good sequence data from public databases here: http://www3.interscience.wiley.com/cgi-bin/fulltext/120748163/sm001.pdf

- Good programs for generating MSAs are Muscle (http://www.drive5.com/muscle or within Jalview through a Web Service), T-Coffee (http://www.tcoffee.org) or other programs in the "T-Coffee-family"from the Notredame group, and MAFFT (http://align.bmr.kyushu-u.ac.jp/mafft/online/server or within Jalview). Expresso, in the T-Coffee family, is very powerful since it will use structural information to improve the MSA. However, it will only work if there are any reasonably close homologs in the PDB. A good, recent, review on how to use the T-Coffee programs is found here: http://www.nature.com/nprot/journal/v6/n11/pdf/nprot.2011.393.pdf

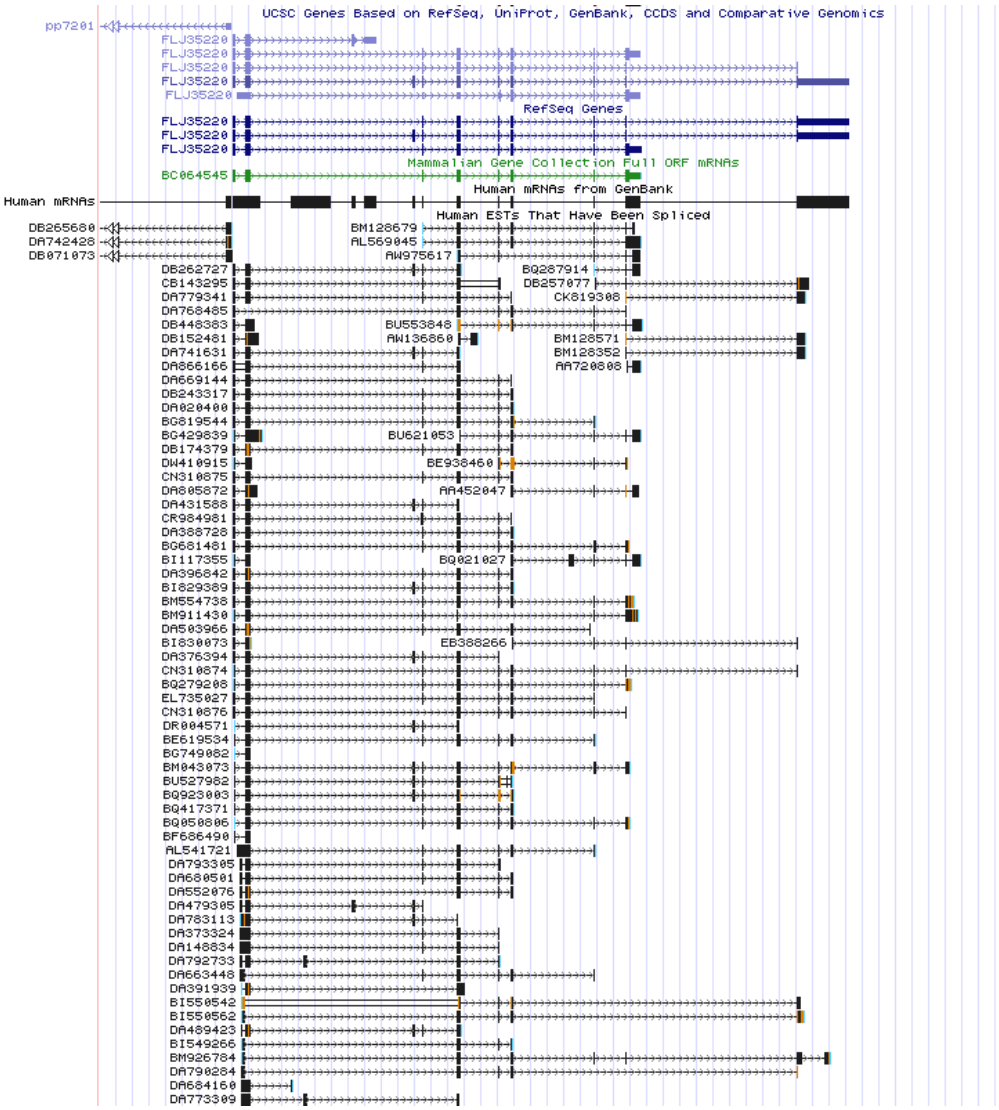
- When you are working with 100s of sequences you might need to download and install the MSA programs on your own server. Experience with scripting languages, Perl, Python or similar will make your work much easier. Example of a task: Make an MSA that "spans the sequence space" of all GPCRs. Take several 1000 GPCR sequences and blast all against all. Make a non-redundant set of sequences with no sequences having more than 95% sequence identity.

- You should of course explore the information about your protein/gene in public databases/tools, for example at NCBI's Protein or protein or gene RefSeq database (http://www.ncbi.nlm.nih.gov/protein), UniProt (http://www.uniprot.org) or Ensembl (http://www.ensembl.org) and in linked databases. ***The perhaps best place to start exploring a protein that is completely new to you is UniProt.***

- ***Of course, read all papers about the protein!!! Or if that is too much, recent review articles.***

- Below is given an example of human protein where nearly all of the sequences in the public databases look strange and clearly are "wrong". However, after some tidying up it can easily be seen to be a highly conserved protein coding gene.

The FLJ35220 (EndoV) locus in the human genome (in UCSC genome browser):



FLJ35220 orthologs (exons 1-6):

```
Mouse       MAHTAAERPPEETLSLWKGEQARLKARVVDRDTEAWQRDPSFSGLQKVGGVDVSFVKGDS
Rat         MAHTATEWPPEETLSLWKGEQARLKARVVDRDTEAWQRDPSFSGLQKVGGVDVSFVKGDS
Dog         MAREAAEKPPEEILSLWKREQAQLKALLVEQDTEAWQRDPAFSGLQRVGGVDLSFVKGDS
Gorilla     MALEAAGGPPEETLSLWKREQARLKAHVVDRDTETWQRDPAFSGLQRVGGVDVSFVKGDS
GuinPig     MAQKAEGQPSEEILLLWKREQARLKALLVDRDTEAWQQDPDFSGLQRVGGVDVSFVKGDS
Orangutan   MALEAAGRPPEETLSLWKREQARLKARVVNWDTEAWQRDPAFSGLQRVGGVDVSFVKGDS
Swine       MARKAARGPQEETLSLWKREQARLKALVVDRDTEAWQRDPAFSGLQRVGGVDVSFVKDDS
Tenrec      MAAPGAGAPPEETLSLWKREQARLKARVVDWDTEAWQRNPDFSGLQRVGGVDVSFVKGDS
Human       MALEAAGGPPEETLSLWKREQARLKAHVVDRDTEAWQRDPAFSGLQRVGGVDVSFVKGDS

Mouse       VRACASLVVLSYPELKVVYEDSRMVGLKAPYVSGFLAFREVPFLVELVQRLQEKEPDLMP
Rat         VRACASLVVLSYPELKVLYEDSRMVGLKAPYVSGFLAFREVPFLVELVQRLQEKEPDLMP
Dog         ASACASLVVLSYPELEVVYEDCSMVNLTAPYMSGFLAFREVPFLVDAVQRLQEKEPHMVP
Gorilla     VRACASLVVLSFPELEVVYEESRMVSLTAPXVSGFLAFREVPFLLELVQQLREKEPGLMP
GuinPig     VTACASLVVLSYPELEVVYEDSRVISLTAPYVSGFLAFREAPFLVDAVHRLQEKEPSLMP
Orangutan   VRACASLVVLSFPELEVVYEESRMVSLTAPYVSGFLAFREVPFLLELVQQLREKEPGLMP
Swine       VSACASLVVLNYPELEVVYEDCRMVSLTAPYVSGYLAFREVPFLVDAVQRLQEKDPQLMP
Tenrec      VNACASLVVLSYPELKVVYEECRMVNLKAPYVSGFLAFREVPFLAEAVQRLQEKEPSLMP
Human       VRACASLVVLSFPELEVVYEESRMVSLTAPYVSGFLAFREVPFLLELVQQLREKEPGLMP

Mouse       QVVLVDGNGVLHQRGFGVACHLGVLTELPCIGVAKKLLQVDGLENNALHKEKIVLLQAGG
Rat         QVLLVDGNGVLHQRGFGVACHLGVLTELPCVGVAKKLLQVEGLENNASHKEKIVLLQAGG
Dog         QVLFVDGNGVLHHRGFGVACHLGILTDLPCIGVAKKLLQVDGLENNAQHKEKIRLLQAEG
Gorilla     QVLLVDGNGVLHHRGFGVACHLGVLTDLPCVGVAKKLLQVDGLENNALHKEKIRLLQTQG
```

```
GuinPig     QVLLVDGNGVLHHQGFGVACHLGVLTGLPCIGVAKKLLQVKGLENNPVHKEKIRLLQAGG
Orangutan   QVLLVDGNGVLHHRGFGVACHLGVLTDLPCVGVAKKLLQVDGLENNALHKEKIRLLQTRG
Swine       QVLFVDGNGVLHHRGFGVACHLGVLTDVPCIGVAKKLLQVDGLENDAAHREKIRLLKAGG
Tenrec      QVLLVDGNGVLHQRGFGVACHLGILTDLLCVGVAKKLLQVDGLEKSDQHKEKVRCLRAAG
Human       QVLLVDGNGVLHHRGFGVACHLGVLTDLPCVGVAKKLLQVDGLENNALHKEKIRLLQTRG
```

This is clearly a conserved and "important" protein in mammals. Still nearly every single mRNA looks unable to give functional protein. If exon 3 is missing, it can be seen from an experimental EndoV structure that a big chunk of the protein core is missing. All mRNAs missing exon 3 are highly unlikely to produce functional protein.

PTMs & domains

When you are fairly confident you have the correct, translated polypeptide in one or several (splice variants) protein sequences it is necessary to consider posttranslational modifications. In many cases (in most cases in for example mammals) the proteins go through post-translational modifications (PTMs) after translation. Some examples are:

- N-terminal Met removal by methionyl-aminopeptidases (MAPs)
  - Only 40% of *E. coli* proteins retain N-terminal Met.
  - Also large fraction of human and yeast proteins processed by MAPs. It is not known which.

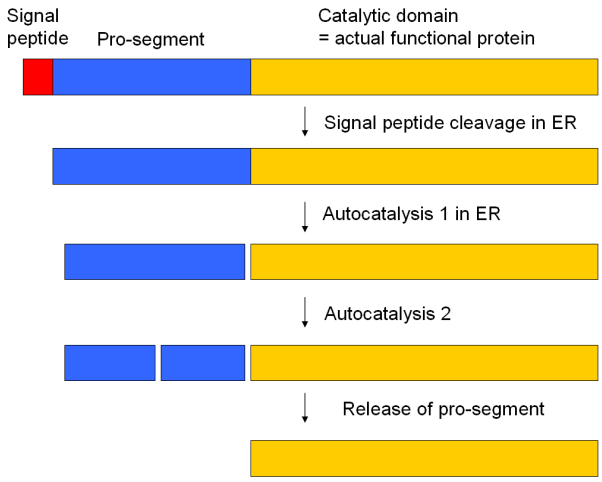
- N-terminal signal peptides are found as pre-sequences that targets the propeptide to the endoplasmic reticulum (ER) and through secretory pathways.
  - Typically 15-30 residues long. A "zip code" for cellular export which is cleaved off in the ER.
  - Examples of predictors for signal peptides are SignalP 4.0. (http://www.cbs.dtu.dk/services/SignalP), Phobious (http://phobius.sbc.su.se), and Signal-BLAST (http://sigpep.services.came.sbg.ac.at/signalblast.html).
  - Read more about signal peptides here: http://www.nature.com/nprot/journal/v2/n4/pdf/nprot.2007.131.pdf

TABLE 1 | Examples of signal peptides[a].

Human α-1-antichymotrypsin precursor (ACT):
MERMLPLLALGLLAAGFCPAVLC↓HPNSPLDEEN...
Escherichia coli class B acid phosphatase precursor:
MRKITQAISAVCLLFALNSSAVALA↓SSPSPLNPGT...
Clostridium perfringens ε-toxin type B precursor:
MKKNLVKSLAIASAVISIYSIVNIVSPTNVIA↓KEISNTVSNE...

[a]Three examples of secretory signal peptides from a eukaryote (Human), a Gram-negative bacterium (*E. coli*) and a Gram-positive bacterium (*Clostridium perfringens*). The cleavage sites are marked by arrows and the hydrophobic regions are underlined. Note that the Gram-positive signal peptide is considerably longer than the Gram-negative, which is slightly longer than the eukaryotic one—these examples have been selected to represent the average length for each organism group.

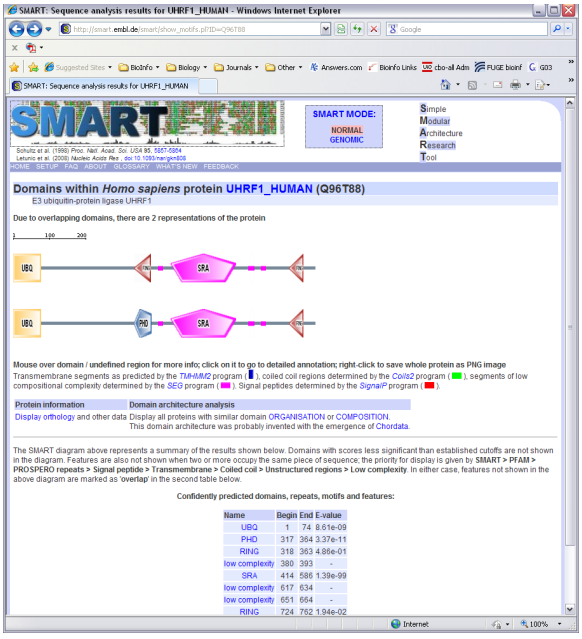
- Many proteins are processed by proteases.
  - For example nine human and a huge amount of bacterial subtilisin-like serine proteases in the proprotein convertase family.



- Many other examples of proteolytic activities. Some processes might not even be known, at least in organisms that are not well studied!

- Other PTMs? Phosphorylation, glycosylation, methylation, ubiquitination, and many more. Also start pondering domains. Is it a single domain protein or can it be divided into several functional domains?

  - Use for example PhosphoSite (http://www.phosphosite.org)
  - or NetPhos (http://www.cbs.dtu.dk/services/NetPhos)
  - or lots of other tools linked from ExPASy (http://www.expasy.ch/tools/#ptm)
  - ELM (http://elm.eu.org)
  - Prosite (http://www.expasy.ch/prosite)
  - InterPro (http://www.ebi.ac.uk/interpro)
  - Pfam (http://pfam.sanger.ac.uk)
  - There are links to many of these resources from sequence databases, for example from UniProt entries.
  - SMART (http://smart.embl.de)



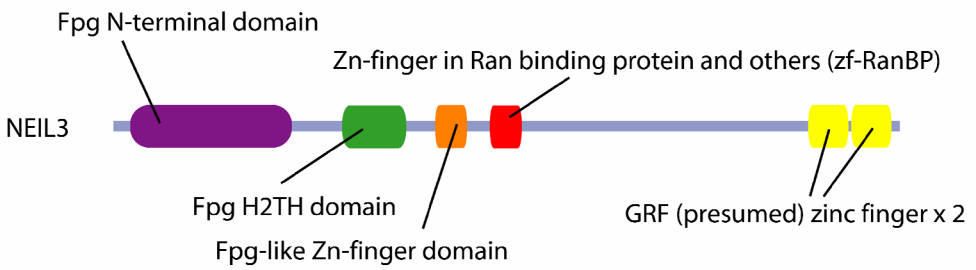
Native structural disorder in proteins

When you have a well-founded idea about the sequence of your processed, functional protein you might consider predicting the 3D structure. ***However, it is extremely important that you check that your protein actually is structured! If your protein is structurally disordered it is meaningless to try to model the structure!***

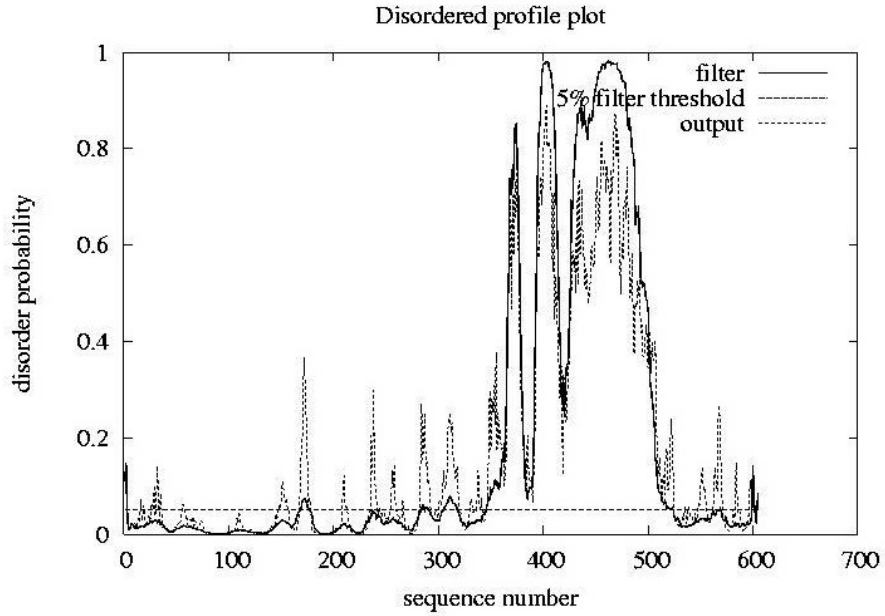
PDB id 1N0Z

- Percentage of proteins in genome with >30 residues predicted to be in structurally disordered segments (Ward *et al.*, J. Mol. Biol. **337**, 635 (2004)):
  - 2% in archaea
  - 4% in eubacteria
  - 33% in eukaryotes
- Good predictors are DISOPRED2 (http://bioinf.cs.ucl.ac.uk/disopred) and some of the PONDR algorithms (http://www.pondr.com with VSL1, VL-XT, etc.) Other disorder predictors are found here: http://www.disprot.org/predictors.php

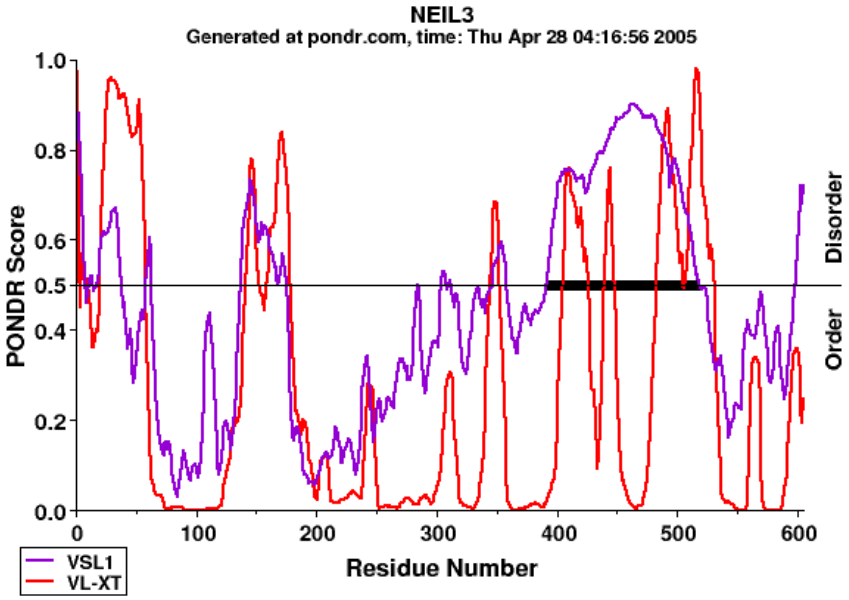
Human NEIL3 protein is a homolog of bacterial Nei/Fpg:

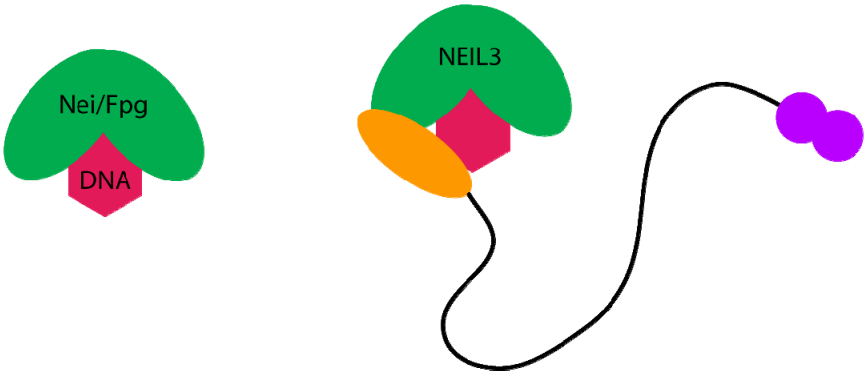


DISOPRED2 for NEIL3:



PONDR (with over prediction?):





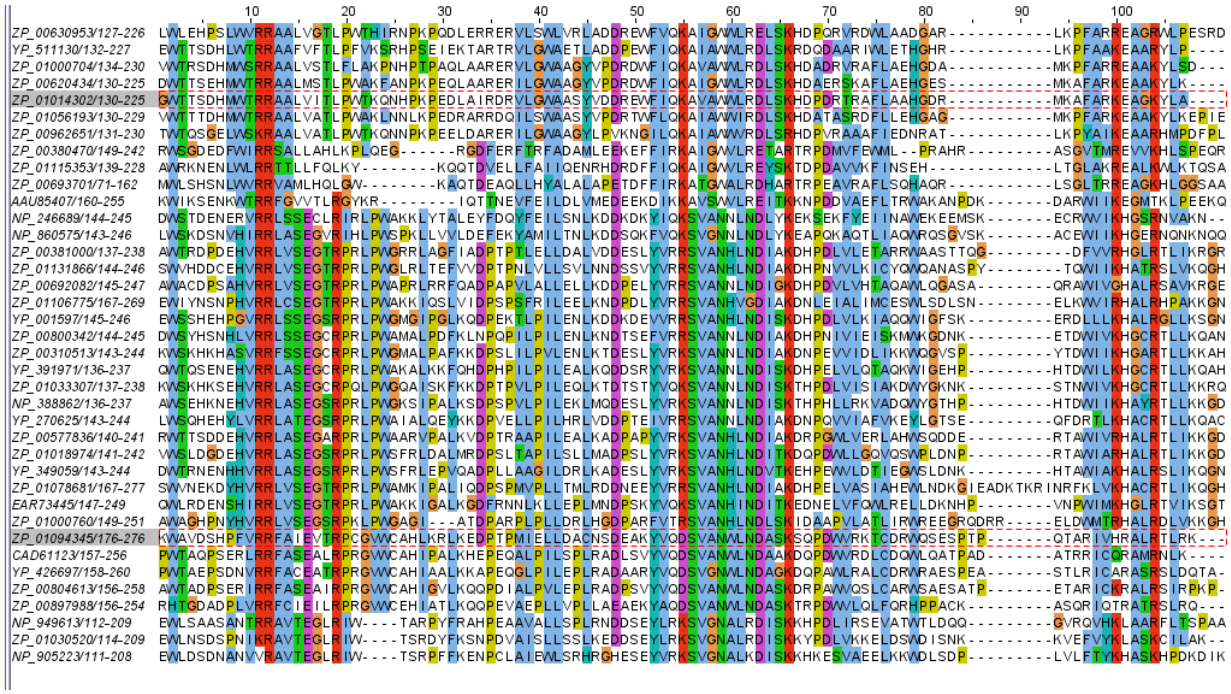
- NEIL3 clearly has 4 N-terminal domains, a long segment (~150 residues) with no regular 3D structure and two similar C-terminal, structured domains. The middle part might be a flexible linker connecting the N- and C-terminal domains. *It is clearly meaningless to model the structure for this part.*

- Read more about structurally disordered proteins here:
  - Ward *et al.*, J. Mol. Biol. **337**, 635 (2004).
  - Dyson & Wrigth, Nat. Rev. Mol. Cell Biol. **6**, 197 (2005).
  - Tompa, Trends Biol. Sci., **27**, 527 (2002).
  - Wright & Dyson, J. Mol. Biol. **293**, 321 (1999).

Predicting 3D structure

If your protein has, or is likely to have, several domains, split the sequence into the putative domains and model the structure for each domain separately. You may combine the domains later if that is possible. "What is a domain" is clearly subjective...

- ***First, try homology modeling.***
  - For each domain, look for structures in the PDB that can be used as templates, for example do blastp in the pdb sequences here: http://blast.ncbi.nlm.nih.gov. Possible templates must have "high enough" sequence identity to your target. "High enough" might be >30%, > 20% or >16% depending on the length of the alignment (longer is better) and the number of indels (fewer is better).
  - If you find no templates: **You cannot do homology modeling!**
  - If you find useful templates, learn to know them... Read the articles that describe them. Choose the one that best suits your needs. Consider resolution, ligands, sequence identity to target etc.
  - Make an alignment that is as good as possible. ***Getting the correct alignment is by far the most important step in homology modeling!*** Now you use the MSA you generated earlier in Jalview with many (50?) homologs of your target and template. Use this MSA to get a good alignment of target and template. The alignment of the two highlighted residues below is much more reliable when it is inferred from the information contained in all the homologous sequences. ***Many sequences good, two sequences pretty bad...!***



- Look at the template in PyMOL or a different visualization program. Make sure all indels are in loops and not inside (long) α-helices or β-strands. If they are, move them to the nearest loop.
- Now you are finally ready to run the homology modeling program!
  - Use Swiss-Model (http://swissmodel.expasy.org)
  - or Modeller (http://salilab.org/modeller)
  - or possibly 3D-JIGSAW (http://bmm.cancerresearchuk.org/~3djigsaw)
- A good tutorial for Swiss-Model is found here: http://www.nature.com/nprot/journal/v4/n1/pdf/nprot.2008.197.pdf
- Maybe there already is a model for your target in a model database? Check for example the Swiss-Model repository (http://swissmodel.expasy.org/repository) or Modbase (http://modbase.compbio.ucsf.edu). It is, however, very likely that you will be able to generate a better model manually!
- Maybe there is something useful at the structural biology Knowledgebase (http://sbkb.org)

- ***If homology modeling is not possible, then try threading/fold recognition or hidden Markov model (HMM) predictions***
  - Phyre (http://www.sbg.bio.ic.ac.uk/~phyre2)
  - GenTHREADER or pGenTHREADER (http://bioinf.cs.ucl.ac.uk/psipred)
  - I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER), with a review/"protocol" paper here: http://www.nature.com/nprot/journal/v5/n4/pdf/nprot.2010.5.pdf
  - SAM-T08 (http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html)
  - Superfamily (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/index.html)
  - If homology modeling does not work ***these methods will very likely not give you a highly reliable model!***
  - These tools might give an indication on remote homology and hint at function.
  - Always use several tools and compare them. If several algorithms give the same result it is more likely to be correct. If all methods give different results you should most likely not trust any of them.

- ***If threading/fold recognition or HMM methods give no reliable hits the only option is ab initio structure prediction.*** Use Robetta (http://robetta.bakerlab.org) or related tools from the David Baker lab (http://depts.washington.edu/bakerpg). *Ab initio* structure prediction is not a mature field and is not likely to give you very reliable models.

Things you can do with the 3D structure

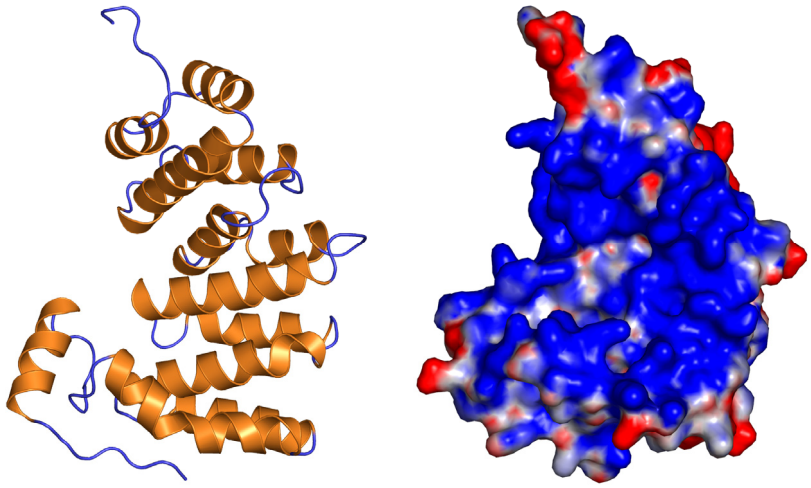
If you were successful in predicting the 3D structure of your target protein or if there already was an experimental structure available in the PDB you may get more biological insights or at least hints by checking out a number of resources.

For experimental structures these sites are great and have lots of links to other resources as well:

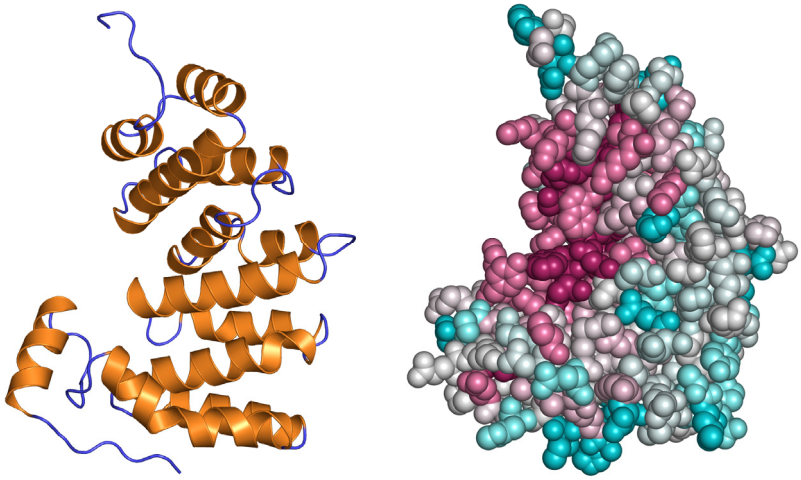
- Protein Data Bank (http://www.pdb.org)
- PDBSum (http://www.ebi.ac.uk/pdbsum)
- NCBI's MMDB (http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml)
- Protein quaternary structures and interactions (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html)

Other stuff to consider:

- Calculate electrostatic potential for example with APBS (http://www.poissonboltzmann.org/apbs). APBS can be run from within PyMOL with the correct PyMOL plugin. PDB2PQR (http://www.poissonboltzmann.org/pdb2pqr) is great for adding H-atoms, correct charges etc. to 3D structures prior to electrostatic potential calculations



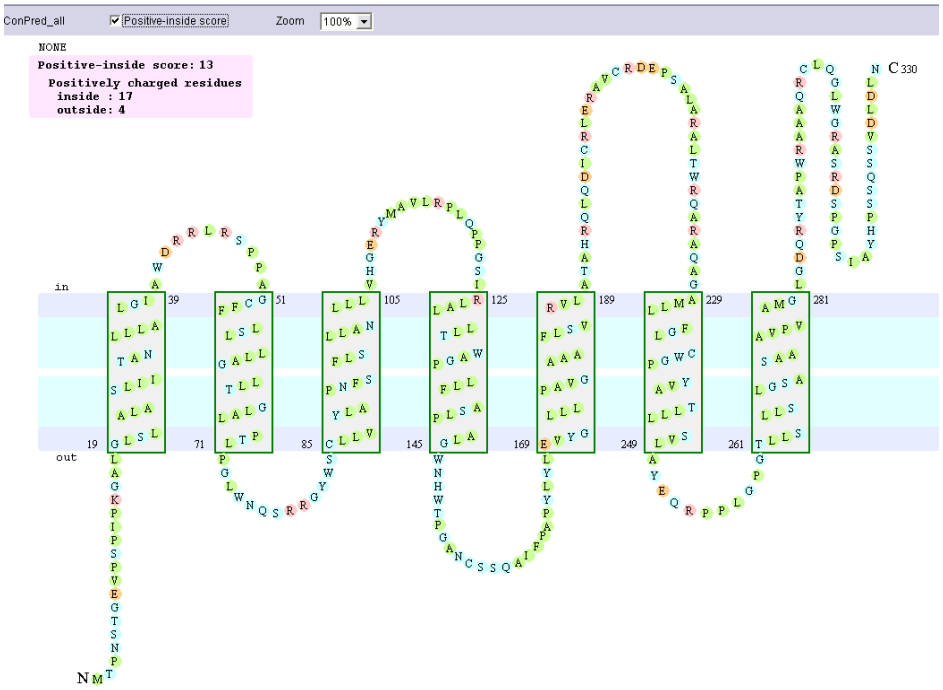
- ConSurf is a great tool for mapping sequence conservation, *i.e.* an MSA, onto a structure (http://consurf.tau.ac.il). It can be used to locate active sites, protein-protein interaction sites etc.



- Use the PyMOL plugins such as AAindex and Protscale (http://www.pymolwiki.org/index.php/AAindex) to look for hydrophobic patches or interesting patterns on the protein surface.
- Suggest residues for site-directed mutagenesis experiments?
  - Patches with several conserved residues on the protein surface?
  - Conserved, hydrophobic residues sticking side-chains out into the solvent?
  - Active sites?
  - Sites for protein-protein interactions or interactions with other macromolecules or metabolites?
- Predicted sites of PTMs like phosphorylation, ubiquitination etc. are much more likely to be correct if they are on the surface.
- A good structure can be used for docking
  - Protein-protein to build complexes?
  - Drug candidates? Inhibitors?
  - For example with HADDOCK: http://www.nature.com/nprot/journal/v5/n5/pdf/nprot.2010.32.pdf
  - Or Glide or other Schrödinger software tools: https://wiki.uio.no/usit/suf/vd/hpc/index.php/Schroedinger_Suite

Other things to consider

- Most structures in the PDB are globular proteins and only a small fraction are membrane proteins. For this reason many predictors and other tools have been optimized for globular proteins. They may not function optimally, and in many cases are useless, for membrane proteins. On the other hand, some tools have been generated specially for membrane proteins, for example transmembrane prediction tools such as Phobius, MEMSAT3 (http://bioinf.cs.ucl.ac.uk/psipred) or ConPred II (http://bioinfo.si.hirosaki-u.ac.jp/~ConPred2)



- A huge amount of links can be found here: http://bioinformatics.ca/links_directory. The question is what to use...
- ***All models must be tested with biological experiments!***
- ***Co-operate with people that know modeling!! Do not work on your own!***