Anyone can edit anything!
PROTEINS

One thousand families for the molecular biologist

Cyrus Chothia

How many families of proteins are there? By putting together the information to be found in papers published over the past few months we can make an initial estimate, and my calculation suggests that the large majority of proteins come from no more than one thousand families.

This work and the proportion of them that have sequences similar to those determined previously.

The six sets of results give a remarkably consistent view: in each case the authors report that close to one-third of the new sequences belong to families that already have members in the sequ-

My personal scientific goal:

“to completely and accurately classify all proteins”
The Pfam Protein Families Database

http://pfam.sanger.ac.uk
T. thermophilus - Ramakrishnan, Cell, 2002
Rfam

A database of non-coding RNA elements
Curating annotation is the bottleneck!
The Masterplan

MCB Wikiproject is awesome!
Community Annotation

• We chose to embrace Wikipedia
  – Saves us implementing a wiki
  – Vast audience
  – Low barrier to entry
  – The Google factor
Family: RyhB (RF00057)

Summary

RyhB RNA

The Rfam group coordinates the annotation of Rfam families in Wikipedia. You can see the Wikipedia page for this family [here](#) . More...

RyhB RNA is a 90 nucleotide non-coding RNA that down-regulates a set of iron-storage and iron-using proteins when iron is limiting; it is itself negatively regulated by the ferric uptake repressor protein, Fur (Ferric uptake regulator). This ncRNA gene was recently identified in a screen and called SraI and was found to be expressed only in stationary phase.\(^2\)

RyhB RNA levels are inversely correlated with mRNA levels for the sdhCDAB operon, encoding succinate dehydrogenase, as well as five other genes previously shown to be positively regulated by Fur by an unknown mechanism. These include two other genes encoding enzymes in the tricarboxylic acid cycle, acnA and furA, two ferroxidase genes, fepA and hfd, and a putative iron storage gene, icsAB.

The figure from Tjaden et al. depicts the secondary structure for the RyhB RNA. The Sm-like protein Hfq binds to the AU-rich unstructured region of RyhB as indicated. Below the secondary structure, the primary sequence of RyhB is shown along with its putative binding interaction to the target mRNA sodB. The start codon for sodB is underlined. RyhB nucleotides that participate in the interaction are in bold.\(^1\)
Who is editing wikipedia pages?
• As of 12th April - 1001 users - 7,645 edits
Vandalism

Riboswitches as antibiotic targets

At least one antimicrobial compound, pyrithiamine, has been shown to act by targeting riboswitches. In this case, cells metabolize pyrithiamine to pyrithiamine pyrophosphate, which was shown to bind and activate the TPP riboswitch. As a consequence, the cell ceases to make TPP. It has also been shown that S-(2-aminoethyl)L-cysteine binds the lysine riboswitch; this could explain why this compound is antimicrobial.

One potential advantage that riboswitches have as an antibiotic target is that many of the riboswitches have multiple instances per genome, where each instance controls an operon containing many genes, many of which are essential. Therefore, in order for bacteria to evolve resistance to the antibiotic by mutations in the riboswitch, all riboswitches must be mutated. However, of course, other mechanisms for resistance may exist, and some -- such as altering the specificity of an exporter to export the drug -- may require fewer mutations.

Riboswitches were parodied in an episode of the simpsons where Homer goes on a tour to eat "Ribwiches"

• User **203.59.66.54** added this line on 3\(^{rd}\) June 2007, removed 17\(^{th}\) June
• Plausible because Ribwiches did appear in Simpsons in 2003
• Most vandalism is reverted in minutes
• Vandalism only made it onto Rfam website once
Getting scientists to edit

- 10 simple rules article in PLoS CB
  - Help scientists understand how to get started
- December 2008 we struck a deal with RNA Biology.
- RNA families track created
- RNA families articles require:
  - Paper describing family
  - Stockholm formatted alignment
  - Wikipedia article
- Paul Gardner acts as Editor
Kinship in the SRP RNA family

Magnus Almqvist, Nils Lerner, Toru Suetake and Christian Zwieb

A family of non-classical pseudoknots in influenza A and B viruses

Alexander P. Gubryszyn and René C. Ossanna

A novel family of plasmid-transferred anti-sense ncRNAs

Sven Fredriksson, Corinna Schmidts, Peter E. Stadler and Ulla Rosner

RsaOG, a new staphylococcal family of highly transcribed non-coding RNA

Arminin Marchais, Chantal Bohn, Philippe Bobeau and Dandan Gamberet
The problem with scientists

Knowledge of subject

Knowledge of Wikipedia

Ignorant Scientists

Wikipedia Pedants
Vanity

• Most scientists gravitate to writing articles about themselves, their groups or projects
• This is an obvious conflict of interest and is discouraged by Wikipedia
• However, writing about a subject you are an expert in is not!

• http://en.wikipedia.org/wiki/Wikipedia:An_article_about_yourself_is_nothing_to_be_proud_of
Example 1

The ChEMBL group is a bioinformatics and cheminformatics group based at the European Bioinformatics Institute (EBI), Hinxton, Cambridgeshire. ChEMBL interests cover Computational Chemical Biology, Chemogenomics, Chemoinformatics, Bioinformatics, Structural Biology, Open Data, Knowledge Management, and Data Integration.

ChEMBLdb

ChEMBLdb is a database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data).

Data Information

The bioactivities are normalised into a uniform set of end-points and units, where possible, and the links between a molecular target and a published assay are tagged with a set of varying confidence levels. The data is abstracted and curated from primary scientific literature and covers a significant fraction of the Structure-Activity Relationships (SAR) and discovery of modern drugs.

At the time of the release of ChEMBL_07 (Oct 2010), ChEMBLdb contained data on 8,078 targets, 602,500 distinct compounds (730,186 compound records) and 2,948,069 activities from 38,204 publications.

Additional data on the clinical progress of compounds is being integrated into ChEMBL at the current time.

Other ChEMBL Databases

Subsets of ChEMBLdb, relating to particular target classes (Kinase SARfari, GPCR SARfari), or disease areas (ChEMBL-NTD), are also available.

These separate data sets, and the entire ChEMBLdb, are available either via download through the website, or via bespoke query interfaces, tailored to the requirements of the scientific communities with a specific interest in these research areas.

ChEMBL Blog

The ChEMBL group keeps their users up to date via their group blog - ChEMBL Blog. The blog informs users about releases, new entries, group activities and future plans. It is also a good place to find out about web walk-throughs and testing of new interfaces.
Example 1

Deletion log

- (del/undel) 21:28, 8 October 2010 Alexf (talk | contribs | block) deleted “ChEMBL” (G11: Unambiguous advertising or promotion) (view/restore)

Restore revisions

To restore the entire page and its history, leave all checkboxes deselected and click Restore. To perform a selective restoration, check the boxes corresponding to the revisions to be restored and click Restore. Selecting a box, then shift selecting another will fill all boxes in between in many browsers. Clicking Reset will clear the comment field and all checkboxes. Please make sure that you are following undeletion policy and that you leave a summary in the comment box.

Comment: [Comment field]

Invert selection | Restore | Reset | Select all

Page history

- (del/undel) (diff) 20:27, 8 October 2010 Jimmy Pitt (talk | contribs | block) (2,612 bytes) (Requesting speedy deletion (CSD G11). (TW))
- (del/undel) (diff) 11:46, 8 October 2010 Chemblgroup (talk | contribs | block) (2,600 bytes)
- (del/undel) (diff) 11:43, 8 October 2010 Chemblgroup (talk | contribs | block) (2,404 bytes)
- (del/undel) (diff) 11:42, 8 October 2010 Chemblgroup (talk | contribs | block) (2,411 bytes)
- (del/undel) (diff) 11:40, 8 October 2010 Chemblgroup (talk | contribs | block) (2,376 bytes)
- (del/undel) (diff) 11:36, 8 October 2010 Chemblgroup (talk | contribs | block) (2,363 bytes)
- (del/undel) (diff) 11:19, 8 October 2010 Chemblgroup (talk | contribs | block) (2,395 bytes)
- (del/undel) (diff) 11:18, 8 October 2010 Chemblgroup (talk | contribs | block) (2,367 bytes)
- (del/undel) (diff) 11:17, 8 October 2010 Chemblgroup (talk | contribs | block) (2,320 bytes) (←Created page with """"ChEMBL"""" is a bioinformatics and cheminformatics group based at the European Bioinformatics Institute (EBI), Hinxton, Camb...)
Example 1
ChEMBL or ChEMBLdb is a manually curated chemical database of bioactive molecules with drug-like properties. It is maintained by the European Bioinformatics Institute (EBI), based on the Wellcome Trust Genome Campus, Hinxton, UK. The database, originally known as SiARlite, was developed by a pharmaceutical company, Galapagos NV. It was acquired for EMBL in 2008 with an award from The Wellcome Trust, resulting in the creation of the ChEMBL chemogenomics group at EBI, led by John Overington.

Scope and access

ChEMBL version 2 was launched in January 2010, including data about 2.4 million activities of 622,824 compounds, including 24,000 natural products. This was obtained from nearly 34,000 publications across twelve medicinal chemistry journals. According to Andreas Bender writing in *Nature Chemical Biology*, ChEMBL’s coverage of available bioactivity data is "the most comprehensive ever seen in a public database." In October 2010 ChEMBL version 8 was launched, with over 2.97 million activities of 636,269 compounds.

ChEMBLdb can be accessed via a web interface or downloaded by File Transfer Protocol. It is formatted in a manner amenable to computerized data mining, and attempts to standardize activities between different publications, to enable comparative analysis. ChEMBL is also integrated into other large-scale chemistry resources, including PubChem and the ChemSpider system of the Royal Society of Chemistry.

Associated resources

In addition to the database, the ChEMBL group have developed tools and resources for data mining. These include Kinase SARfari, an integrated chemogenomics workbench focussed on kinases. The system incorporates and links sequence, structure, compounds and screening data. GPCR SARfari is a similar workbench focussed on GPCRs and ChEMBL-Neglected Tropical Disease (ChEMBL-NTD) is a repository for Open Access primary screening and medicinal chemistry data directed at endemic tropical diseases of the developing regions of the Africa, Asia, and the Americas. The primary purpose of ChEMBL-NTD is to provide a freely accessible and permanent archive and distribution centre for deposited data.

See also

- ChEBI
- DrugBank

References


External links

- ChEMBLdb
- Kinase SARfari
- ChEMBL-Neglected Tropical Disease Archive
- GPCR SARfari
- The ChEMBL-log Open data and drug discovery blog run by the ChEMBL team.
Example 2

• User Alexm1313 create page on Frozen Ark on 19th July.

• Page rapidly grows
Example 2

Content dispute

I have tried to engage Alexm1313 (talk • contribs) in a discussion about content he has added to this article, specifically material in the lede regarding the prospects of future extinctions and the supposed efficacy of this project. As I pointed out on his talk page:

• There is no need for a discussion of rates of extinction on this page, as that topic is thoroughly covered elsewhere in Wikipedia, and covering it here simply means that the information must be maintained in multiple places.
• There is no evidence that the information will be extremely important to future scientists. Clearly, there is a hope that this will be the case (else the project is useless), but there is no evidence to support this claim.
• Stylistically, one would never say "animals are facing a huge amount of extinction".
• The expectation that 1/4 of the world's species will go extinct over the next 30 years is uncited, and is unnecessary in this article. By linking to the article on extinction, we can provide a link to the necessary information without duplicating it (or providing incorrect information) in this article. Duplication of information across multiple articles is called content forking and is discouraged.

However I have been unable to elicit a response from this user other than a continued re-insertion of the disputed material. Other opinions are welcome. WikiDan61, ReadMe!! 17:28, 19 July 2010 (UTC)

Lots of discussions later …

I am not interested in people being horrible, and I want no affiliation to that page. Poor newbies who have to put up with that. Wasted hours of my work time for something very simple. That is it for me, I am not interested in that kind of behaviour.
Attribution

• Many scientists are worried about lack of attribution

SmY RNA

A C-class article from Wikipedia, the free encyclopedia

SmY ribonucleic acids (SmY RNAs) are a family of small nuclear RNAs found in some species of nematode worms. They are thought to be involved in mRNA trans-splicing.

SmY RNAs are about 70-90 nucleotides long and share a common secondary structure, with two stem-loops flanking a consensus binding site for Sm protein.\(^2\)\(^3\) Sm protein is a shared component of spliceosomal snRNPs.

SmY RNAs have been found in nematodes of class Chromadorea, which includes the most commonly studied nematodes (such as Caenorhabditis, Pristionchus, and Ascaris), but not in the more distantly related Trichinella spiralis in class Dorylaimia. The number of SmY genes in each species varies, with most Caenorhabditis and Pristionchus species having 10-26 related paralogous copies, while other nematodes have 1-5.\(^1\)
Attribution

• Scientists still want to know what do I get out of editing Wikipedia
• Do tenure committees consider this, etc…

• This rather misses the point