HHS Public Access

Author manuscript

Methods Mol Biol. Author manuscript; available in PMC 2015 October 12.

Published in final edited form as:

Methods Mol Biol. 2014 ; 1140: 53–59. doi:10.1007/978-1-4939-0354-2_4.

Selecting targets from eukaryotic parasites for structural genomics and drug discovery

Isabelle Q. H. Phan[1,2], **Robin Stacy**[1,2], and **Peter J. Myler**[1,2,3,4]

[1]Seattle Structural Genomics Center for Infectious Disease, Seattle, Washington

[2]Seattle Biomedical Research Institute, Seattle, Washington

[3]Department of Global Health, University of Washington, Seattle, Washington

[4]Department of Medical Education and Biomedical Informatics, University of Washington

Abstract

The selection of targets is the first step for any structural genomics project. The application of structural genomics approaches to drug discovery also starts with the selection of targets. Here, three protocols are described that were developed to select targets from eukaryotic pathogens. These protocols could also be applied to other drug discovery projects.

Keywords

Target selection; Structural genomics; Eukaryotes; Pathogens; Drug targets; Bioinformatics

1 Introduction

The success of the methods developed in early structural genomics projects led to the application of the approach to specific scientific problems. Early projects focused on the exploration of protein structure and selected targets for that purpose [1, 2]. More recently, there has been a transition to applying structural genomics methods to structure-aided drug discovery. The Seattle Structural Genomics Center for Infectious Disease (SSGCID) has focused on providing experimentally determined structures that serve as an initial step in the development process of structure-based drugs, vaccines, and diagnostics for infectious diseases [3, 4].

The SSGCID has focused its structure determination efforts on human pathogens, including bacterial, eukaryotic, and viral organisms from the National Institute of Allergy and Infectious Diseases (NIAID) list of biodefense organisms and those causing emerging and re-emerging diseases. Here, we describe three protocols for selecting targets as applied to seven eukaryotic pathogens: *Babesia bovis*, *Coccidioides immitis*, *Cryptosporidium parvum*, *Encephalitozoon cuniculi*, *Entamoeba histolytica*, *Giardia lamblia*, and *Toxoplasma gondii*.

CORRESPONDING AUTHOR: Peter J. Myler, peter.myler@seattlebiomed.org.

The first protocol involves identification of potential drug targets in the seven eukaryotic species listed above. The approach involves searching for proteins with sequence similarity (>50 % over >75 % of their length) with protein targets in the DrugBank database [5]. DrugBank represents a comprehensive and publicly available resource that combines detailed drug (i.e., chemical) data with comprehensive target (i.e., protein) information [5]. The database contains over 6,700 drug entries and over 4,000 non-redundant protein sequences that are linked to these drug entries. Selecting proteins with sequence similarity to known drug targets substantially increases the likelihood that selected proteins are "druggable." In addition, knowledge of chemical ligands (e.g., the drugs that act against their DrugBank homologues) that are likely to bind these proteins should increase their success in traversing the structure determination pipeline and provide ligands for co-crystallization. Determination of their three-dimensional structures will facilitate basic biomedical research by significantly shortening the time needed for development of novel chemotherapeutic agents. The application of this protocol yielded a total of 679 targets in the seven chosen eukaryotic pathogens.

The second protocol focuses on identifying representatives of hand-selected drug candidates. In this case, drug targets were obtained through a literature survey, discussions, and communications with pharmaceutical and academic researchers. A total of 93 targets representing 42 protein families from 32 organisms were collected. Orthologs were identified in the above seven eukaryotic pathogens using OrthoMCL clustering. OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences [6]. It provides not only groups shared by two or more species/genomes but also groups representing species-specific gene expansion families. This protocol yielded a total of 65 targets in the seven chosen eukaryotic pathogens.

The third protocol uses the TDRtargets public repository [7]. The TDRtargets project has collected diverse information relevant to drug target identification for a variety of important human pathogens [7] and provides a website where researchers can look for information on targets of interest. In addition, by using the TDRtargets database tools, researchers can quickly prioritize genes of interest by running simple queries (such as looking for small enzymes or proteins with high-quality structural models), assigning numerical weights to each query (in the history page), and combining these results to produce a ranked list of candidate targets. This protocol yielded a total of 614 targets in the eukaryotic pathogens of the genera *Babesia*, *Brugia*, *Cryptosporidium*, *Leishmania*, *and Trypanosoma*.

The methods in this chapter describe the following steps of the target selection strategy: (1) creating the reference genome sequence dataset, (2) selecting candidate targets, and (3) filtering out sequences not conducive to structural genomics approaches. Data management is a major component of target selection in structural genomics; however, it is outside the scope of this chapter.

2 Materials

1. Computer running the UNIX operating system, Internet connection (*see* Note 1).

2. Installed bioinformatics software: NCBI blast [8], OrthoMCL [6], Phobius [9], or TMHMM [10] (*see* Note 2).

3. Proficiency in a scripting language such as Python or Perl for parsing and combining the results of the target selection steps. A workable alternative that has been tested is the cloud-based database service SQLShare [11].

3 Methods

3.1 Create the Reference Genome Sequence Dataset

1. Select a representative strain for each genus (*see* Note 3).

2. Download CDS (i.e., DNA) and protein sequences in fasta format from EupathDB [12], which stores all organisms described in the introduction, except for *Coccidioides*, for which the sequences were downloaded from the Broad Institute (*see* Note 4).

3.2 Select Candidate Targets

3.2.1 DrugBank Homologues

1. Download target protein sequences from DrugBank in fasta format.

2. Perform a sequence similarity search using BlastP of these reference sequences against DrugBank, and keep the hits with at least 50 % similarity over 75 % of their length (*see* Note 5).

3. The remaining protein sequences were Jaccard clustered [13] to remove paralogs that shared >75 % similarity over 75 % of their length.

3.2.2 Representatives of Known Drug Targets

1. Obtain protein sequences of nominated drug candidates from the relevant repository if a database identifier is provided or search the UniProt database (*see* Note 6).

---

[1]The UNIX operating system is required for installing the stand-alone versions of the sequence clustering and transmembrane prediction tools that we describe. There are many cross-platform alternatives; for example, the Jaccard algorithm is widely used for clustering sequences and is available in a variety of languages, including R and Perl, and transmembrane predictions can be obtained by querying the TMHMM web service http://www.cbs.dtu.dk/ws/ws.php?entry=TMHMM.

[2]The current version of OrthoMCL (v.2.0) requires a relational database. The version we used (v.1.4) does not require a relational database; it is still available for download but is no longer supported.

[3]The choice may be restricted by the availability of fully annotated genomes, as many genomes are first published as unassembled contigs. Beware that genome sequences in EupathDB are continuously updated between releases and sequence quality can vary widely depending on the organism.

[4]Different repositories will store different gene predictions and annotations, or different versions, of the same genome. Beware of inconsistent CDS (ORF) and protein sequences, due, for example, to frameshifts or truncations, duplicate sequences, missing start or stop codons, as well as use of an asterisk at the end of protein sequences and non-ASCII characters in the annotation, which may affect downstream sequence analysis. In EupathDB, proteins that contain asterisks within the sequence are likely to indicate a pseudogene and can be discarded.

[5]Choose the blast+ tabular output with the option “-outfmt ppos” to obtain the percent similarity (or conservation) as the percentage of positive-scoring matches. Several new options for customizing the tabular output format were introduced in blast+ version 2.2.28, including the option ‘stitle’ to display the product description.

[6]Due to the non-standardized nature of protein annotation, using a combination of as many search terms as possible, such as gene, product, function, and organism name, and checking position-specific annotations (such as active sites) will increase the likelihood of finding the correct sequence [18].

2. Combine sequences from the first step with reference protein sequences obtained in Subheading 3.1, and perform all-against-all BlastP, followed by OrthoMCL clustering using a Markov inflation index of 1.2 (*see* Note 7).

3. Select sequences from the reference genomes that cluster with the original nominated sequences from **step 1**.

3.2.3 TDR Targets

1. Query database using the TDRtargets "search for targets" web form http://tdrtargets.org/search (*see* Note 8).

2. Download results. In the horizontal menu on top of the page, click on "my queries." The results are listed under "My target queries," click on the "Export" link, and export using the default format.

3. Download sequences from the Source Database (*see* Note 9).

3.3 Remove Targets Not Conducive to Structural Genomics

1. Remove targets containing introns unless cDNA is available (*see* Note 10).

2. Screen proteins with known structure or those selected by other structural genomics centers to remove targets showing greater than 95 % conservation and 95 % coverage to targets in the Structural Biology Target Registration Database (TargetTrack, formerly TargetDB [14]) and sequences in the Protein Data Bank [15] by performing a BlastP search against these two databases (*see* Note 11).

3. Remove targets that contain transmembrane domains predicted by TMHMM or Phobius (*see* Note 12), except for N-terminal signal sequences, which are removed before PCR amplification.

4. Remove targets that are longer than 750 amino acid residues in length and have a cysteine content greater than 10 (*see* Note 13). Those criteria are "rules of thumb," but it is known that limiting the number of cysteine residues decreases the likelihood of protein aggregation.

---

[7]The OrthoMCL (v.1.4) Markov Inflation Index was reduced to 1.2 from its default value of 1.5 in order to obtain larger clusters of more distant relatives.

[8]The original query for the SSGCID target selection was published as http://tdrtargets.org/published/browse/t/390.

[9]Click on one of the links in the gene_name column of the exported spreadsheet to check the Source Database. Beware that the Source Database may differ for each organism.

[10]Information on introns and exons is stored differently, depending on which database the sequence came from. In EupathDB, the number of exons conveniently appears as a gene attribute in the "Select Column" menu at the top of the search results table. In GenBank records, the exon locations appear in the CDS section of the features.

[11]The TargetTrack and PDB databases are updated weekly. Including the target status in the TargetTrack fasta header allows the recovery of targets that have been marked as "work stopped" and are thus no longer pursued by the depositor. However, this requires building a custom fasta file from the XML format as the status is not included in any of the target protein fasta files provided on the TargetTrack website.

[12]TMHMM and Phobius predictions are limited to helical transmembrane domains. There are no established predictors yet for transmembrane beta-barrels. This is relevant insofar as transmembrane beta-barrels are present in the mitochondria of Eukaryotes.

[13]This is a trivial computing task; however, these values can also be computed using online sequence analysis tools such as ProtParam (http://web.expasy.org/protparam/). In SQLShare, assuming that the protein sequence is in upper-case, the number of cysteines is easily obtained via the SQL statement: SELECT len([sequence])-len(replace([sequence],'C',″)) AS cysteines.

Established high-throughput centers will run these steps routinely using customized automated software pipelines. The SSGCID uses the Ergatis workflow management system that executes jobs in parallel on a computer cluster [16]. Another popular tool is the Galaxy platform [17].

Acknowledgments

The SSGCID has been funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract Nos.: HHSN272201200025C and HHSN272200700057C.

References

1. Brenner SE. Target selection for structural genomics. Nat Struct Biol. 2000; 7(Suppl):967–969.10.1038/80747 [PubMed: 11104002]
2. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B. Automatic target selection for structural genomics on eukaryotes. Proteins. 2004; 56(2):188–200.10.1002/prot.20012 [PubMed: 15211504]
3. Myler PJ, Stacy R, Stewart L, et al. The Seattle structural genomics center for infectious disease (SSGCID). Infect Disord Drug Targets. 2009; 9(5):493–506. [PubMed: 19594426]
4. Stacy R, Begley DW, Phan I, et al. Structural genomics of infectious disease drug targets: the SSGCID. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2011; 67(Pt 9):979–984.10.1107/S1744309111029204
5. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006; 34(Database issue):D668–D672.10.1093/nar/gkj067 [PubMed: 16381955]
6. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13(9):2178–2189.10.1101/gr.1224503 [PubMed: 12952885]
7. Aguero F, Al-Lazikani B, Aslett M, et al. Genomic-scale prioritization of drug targets: the TDR Targets database. Nat Rev Drug Discov. 2008; 7(11):900–907.10.1038/nrd2684 [PubMed: 18927591]
8. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421.10.1186/1471-2105-10-421 [PubMed: 20003500]
9. Kall L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004; 338(5):1027–1036.10.1016/j.jmb.2004.03.016 [PubMed: 15111065]
10. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001; 305(3):567–580.10.1006/jmbi.2000.4315 [PubMed: 11152613]
11. Howe, B.; Cole, G.; Souroush, E., et al. Database-as-a-service for long-tail science. In: Bayard, Judith; Cushing, JF.; Bowers, Shawn, editors. Lect Notes Comput Sci; Proceedings of the 23rd international conference on Scientific and statistical database management; Portland. 2011; Springer-Verlag; p. 480-489.
12. Aurrecoechea C, Brestelli J, Brunk BP, et al. EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res. 2010; 38(Database issue):D415–D419.10.1093/nar/gkp941 [PubMed: 19914931]
13. Jain, AK.; Dubes, RC. Prentice Hall advanced reference series. Prentice Hall; Englewood Cliffs, NJ: 1988. Algorithms for clustering data.
14. Chen L, Oughtred R, Berman HM, Westbrook J. TargetDB: a target registration database for structural genomics projects. Bioinformatics. 2004; 20(16):2860–2862.10.1093/bioinformatics/bth300 [PubMed: 15130928]
15. Rose PW, Bi C, Bluhm WF, et al. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res. 2013; 41(D1):D475–D482.10.1093/nar/gks1200 [PubMed: 23193259]

16. Orvis J, Crabtree J, Galens K, et al. Ergatis: a web interface and scalable software system for bioinformatics workflows. Bioinformatics. 2010; 26(12):1488–1492.10.1093/bioinformatics/btq167 [PubMed: 20413634]
17. Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005; 15(10):1451–1455.10.1101/gr.4086505 [PubMed: 16169926]
18. Hinz U. From protein sequences to 3D-structures and beyond: the example of the UniProt knowledgebase. Cell Mol Life Sci. 2010; 67(7):1049–1064.10.1007/s00018-009-0229-6 [PubMed: 20043185]