J Mol Model (2009) 15:695–699
DOI 10.1007/s00894-008-0353-5

ORIGINAL PAPER

Meta-basic estimates the size of druggable human genome

Dariusz Plewczynski · Leszek Rychlewski

Received: 10 April 2008 / Accepted: 10 July 2008 / Published online: 29 July 2008

Abstract We present here the estimation of the upper limit of the number of molecular targets in the human genome that represent an opportunity for further therapeutic treatment. We select around ~6300 human proteins that are similar to sequences of known protein targets collected from DrugBank database. Our bioinformatics study estimates the size of 'druggable' human genome to be around 20% of human proteome, i.e. the number of the possible protein targets for small-molecule drug design in medicinal chemistry. We do not take into account any toxicity prediction, the three-dimensional characteristics of the active site in the predicted 'druggable' protein families, or detailed chemical analysis of known inhibitors/drugs. Instead we rely on remote homology detection method Meta-BASIC, which is based on sequence and structural similarity. The prepared dataset of all predicted protein targets from human genome presents the unique opportunity for developing and benchmarking various in silico chemo/bio-informatics methods in the context of the virtual high throughput screening.

Keywords Compound identification · DrugBank database · Druggable human genome · Human drug targets · MDL drug data report · Protein target specificity · Virtual high-throughput screening

D. Plewczynski (✉)
Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw,
Pawinskiego 5a,
02–106 Warsaw, Poland
e-mail: D.Plewczynski@icm.edu.pl

L. Rychlewski
Bioinformatics Laboratory, BioInfoBank Institute,
Poznan, Poland

Introduction

The 'druggable' proteins are those that are able to bind drug-like small chemical molecules. Some of those 'druggable' proteins can have altered their biological functions through complex formation, but such modulations may not provide any therapeutically significant effect. Therefore the actual limit of the size of druggable genome can be lower, and only real life experiments can confirm each selected protein as the drug target. Nevertheless in silico whole genome analysis of similarity between known protein targets and human proteins provide the excellent starting point for further selection of real drug targets in the context of a specific disease.

The druggability of a novel protein can be obtained from its experimentally studied homologue. We are following here the paradigm: "one gene, one target", i.e. we treat single genes as druggable without considering the whole biological complexes (multimers). This similarity can be estimated using sequence or structural information. Reliable sequence similarity search tools such as Fasta [1–4] or Blast [5–7] are frequently not powerful enough to detect homology unambiguously. On the other hand threading methods are able to find remote homologues, when one of the proteins to be aligned has a known three-dimensional structure. Those structural methods were developed to detect protein analogues, i.e. structurally similar proteins, yet without evolutionary relationship. Most of the predictions from threading approaches were later found to be homology based by more advanced sequence comparison methods, such as PSI-Blast [5, 6, 8, 9].

On the other hand, structural genomics projects are still unable to assign fold for complete proteomes. Experimentally determined structures remain unreachable for many important prokaryotic and eukaryotic proteins. Therefore,

biologists seek cheap and rapid computational methods for fold assignment. Those prediction algorithms provide valuable structural information for many targets, which can guide the experimental work on selected protein targets [10]. The further development of those algorithms has been boosted in the last years by community-wide assessment experiments in CASP, i.e. critical assessment of techniques for protein structure prediction [11, 12]. Our computational protocol, i.e. *Meta-BASIC*, is focused on the extracting remote sequence similarity between proteins in the high throughput scale. The integrated service incorporates confidence values both at the protein and residue level, and internal quality checks. The method is based on the protein family analysis, predicted domain boundaries and other computational approaches. The output can be easily used for structure based functional annotation. The Meta-BASIC http://basic.bioinfo.pl [13] uses bilaterally amplified sequence information comparison tool, that combines the achievements in sequence profile-based strategies with secondary structure predictions to generate fast and reliable predictions using meta profile alignment methods. Meta-BASIC is proven to outperform many individual servers, including fold recognition servers, and it can compete even with meta-predictors. In addition, Meta-BASIC enables detection of very distant relationships even if the tertiary structure for the reference protein is not known, and has a high-throughput capability [13].

The number of molecular targets in the human genome representing an opportunity for therapeutic treatment was up to now estimated by only sequence homology. Sequence based methods apply extensive similarity search of homologous sequences for those that are known to be protein targets from various experiments [14]. Their estimates are similar and depend on the size of the dataset used as gold-standard (i.e. the number of known 'druggable' proteins). Early work of Drews et al. (1997) [15] presents 483 known targets and estimates the number of disease-modifying genes to be around 5,000–10,000 [15, 16]. Most of known drugs compete for a binding site in a protein with endogenous small molecules; therefore they have to bind to the target with higher potency.

The first high throughput estimate of the size of druggable human genome was done by comparing sequences of known therapeutically relevant targets with sequences from large genome databases. Using only sequence similarity Hopkins and Groom [17] estimated the number of disease related genes to be around 10% of the whole genome. They identified 399 non-redundant molecular targets that bind small chemical molecules with affinity below 10 μM. Authors hypothesized that if one member of a given gene family was able to bind a drug-like molecule; other proteins from the family would be also able to bind a similar compound. Following this paradigm they concluded that 14% of proteins in proteome could be predicted as 'druggable' targets. Their work was later repeated by Russ and Lampel [18]. They estimated using sequence similarity the number of druggable genes to be around 3,000. Authors collected 2,917 druggable genes from Ensembl [19, 20] and 1,942 from Consensus Chemical Database Service (CCDS). They used updated version of InterPro database [21, 22], together with Pfam protein domain classification [23–26]. The authors collected 2917 druggable genes from Ensembl and 1942 from CCDS.

The present study extend those results by including not only sequence similarity, but also more remote similarity between proteins taking into account also the secondary structure patterns. Therefore it includes structural similarity between proteins, thus extending the size of the set of potential druggable targets.

Methods and results

The most important point in the prediction of protein 'druggability' is the selection of the database of protein targets. We use here the massive sequence data from http://redpoll.pharmacy.ualberta.ca/drugbank/ covering all known protein targets. The selected source of medicinal data provide solid ground for the estimation of the number of proteins that can be used in pharmaceutical applications and further analysis of sequence and structural details of possible interactions. This comprehensive resource is ideal for in silico drug discovery by linking chemoinformatics data with bioinformatics resources [27]. It now contains approximately 4,300 drug entries: over 1,000 FDA-approved small molecule drugs, 113 FDA-approved biotech, i.e. protein/peptide drugs, 62 nutraceuticals and over 3,000 experimental drugs. Over 6,000 protein sequences are linked as drug target sequences to these drug entries.

The present study extends the sequence similarity based estimation of the size of druggable human genome by including also structural information. The Meta-BASIC approach is based on four crucial components. The first one presents a novel sequence profile-based method. Profile methods, including PSI-Blast, set the standard in the field as very accurate predictors of remote links between proteins. High-scoring PSI-Blast hits are essentially correct and biologically meaningful. In addition, a skilful PSI-Blast user is able to pick a few non-trivial homologues by careful analysis of hits in the twilight-zone. However, many interesting but very remote homologues still remain undetected at the sequence level. The second one uses predicted local structure information. Adding structural information to a sequence profile helps to find those homologues that diverged beyond recognition sequence-wise, but remain structurally similar. In contrast to meta-profiles tools, many conventional threading algorithms use

experimental global structure to score similarity. Therefore a protein of interest should have a structure of its homologue determined as a pre-requisite for correct prediction. The Meta-BASIC is free from that requirement and can find links between proteins of unknown structure. In addition, parting with the global threading allows for a faster algorithm and higher throughput. The third one not only combines sequence profile with secondary structure profile to form what we call "meta"-profile, but also utilizes several scoring systems and alignment algorithms. Averaging between the results obtained by slightly different approaches helps to boost the accuracy. Those three components provide a high-throughput capability since it is a stand-alone program in contrast to most meta-servers that collect predictions from several remotely located servers. The fourth component includes protocols for communication with remote servers. It collects predictions from several external servers, publicly available databases and services.

In the present study we prepared 38,171 human proteins from HPI (the UniProtKB/Swiss-Prot Human Proteome Initiative at http://www.expasy.org/ sprot/hpi/hpi_desc.html) proteome database. This constructed the proteome of interest from which they selected those proteins that had direct BLAST hit (e-value equal to 0.0) to protein sequences from DrugBank database (http://redpoll.pharmacy.ualberta.ca/drugbank/index.html). The first step of analysis was to cluster the whole set of redundant proteins by CD-HIT [28] to get 6,414 proteins. Next, the human database was searched by using Meta-BASIC on this potential target database. The most interesting group of human proteins is the set with high scoring hits to DrugBank and PDB proteins. Those preliminary results suggest that ~6,300 redundant human sequences have conservative hits to DrugBank target sequences. This estimates the size of 'druggable' human genome to be around 20% of whole human proteome. It is a very rough estimate (we did not calculate any redundancy of sequences in DrugBank, or the redundancy of sequences in EBI/HPI human protein database). On the other hand the non-redundant subset of over 3,000 sequences from DrugBank protein targets hade direct Meta-BASIC hits (e-value less than 0) to human proteins. The non-redundant selection was performed using CD-HIT with cut-off 1.0. In total, around 2,825 proteins from DrugBank (6,414 non-redundant subset) have direct sequence similarity to proteins from PDB90 database and human proteome.

Our recent studies confirm results by Hopkins and Groom [17] done by comparing sequences of known therapeutically relevant targets with sequences from large genome databases. In their approach sequences of those drug-binding domains were used for identification of 130 families of InterPro database of domains [21, 22] that contain at least one known drug target. Our predicted set of druggable proteins has similar percentages of known, large protein families. We have also found GPCRs (G-protein-coupled receptor) proteins, protein kinases (with serine/threonine/tyrosine modifications), zinc metallopeptidases, serine proteases, nuclear hormone receptors and phosphodiesterases. We assume, that if at least one member of a given gene family was able to bind a drug-like molecule; other proteins from the family would also be able to bind a similar compound. The two most populated protein families are GPCRs and protein kinases. The protease group was found to be the third most important one. The classification of predicted protein targets by their biochemical characteristics reported similar results to previous studies. Most protein targets appear to be enzymes (~50%) and GPCR proteins (~30%), and ion channels ~10%. The rest of medium or small protein groups are transporters, nuclear hormone and other receptors, integrins and DNA.

This classification is similar to work by Russ and Lampel [18] and results published by Plewczynski [14]. This work reported that around 5,800 human sequences have conservative BLAST [6] hits to sequences from DrugBank database of protein targets. Their estimate of the 'druggable' human genome size is equal to around 15%. These estimates did not take into account any redundancies of sequences in DrugBank, or the redundancy of sequences in EBI/HPI human protein database. No significantly populated new classes of proteins were identified when Meta-BASIC was applied to search for remote sequence similarity, compared to the results obtained from those sequence comparison. Therefore further and more refined prediction of druggability have to be done by structure based approaches, when unexpected structural similarity not detectable by sequence methods will reveal new functional associations between protein of human proteome. We hypothesize that global comparison of proteins structures, or more focused local comparison of active sites structures can reveal new functionally similar pairs of proteins. Therefore such an approach is likely to provide new protein families similar to those that are druggable, yet are not detectable at the global sequence level.

Summary

The whole 'druggable' human genome is the target for currently developed small chemical molecules in typical drug design procedure. We did not consider any other approaches for affecting the diseases by applying other biomolecules (for example peptides). It is estimated that only four novel targets are launched on the marked each year with similar number of new chemical entities (NCEs).

In 2002 only 120 proteins were known with marketed drugs [17]. Therefore, even a small number of targets is enough to successfully build pharmaceutical industry. Oral bioavailability of molecules follows the rule-of-five first proposed by Lipinski in 1997. Yet new protein drugs, antibodies, DNA vaccines, virus therapies could expand the presently known range of potential drug targets. In this article the paradigm: "one gene, one target" is closely followed, i.e. we treat single genes/proteins as 'druggable' without considering the whole biological complexes (multimers). This assumption is of course questionable as sometimes only multimers are medicinally active. However, in our approach we neglect this information even if it is available from PDB database or literature. Moreover, many successful drugs have more than one molecular target and their utility in therapy is determined by the balance between their actions (not their absolute specificity). This is also neglected in the presented approach.

Those theoretical methods strongly depend on the available experimental data. As known from previous studies the 'druggability' of the protein is typically defined as its ability to bind a small chemical molecule. Therefore when a set of proteins is available that are experimentally confirmed as 'druggable' ones, it is possible to propose new proteins that can be inhibited in a similar way. Thus performing sequence similarity search on this set is usually enough to find new potent candidates for 'druggable' proteins. Those proteins are likely to share similar three dimensional fold, therefore they are able to bind drug-like small molecules. We stress here, that it does not guarantee the linkage to particular diseases. Those candidates have to be verified by detailed structural analysis, other in silico virtual high-throughput methods, and ultimately by experimental means in order to confirm that activities of these proteins can actually be inhibited by any of known drugs. Nevertheless, the final set of protein targets can be mapped onto whole human proteome providing the reliable estimate of the size of human 'druggable' genome. It should be stressed that these proteins are of crucial importance if one is searching for drug targets for a given disease or one is performing off-target specificity analyses for a small chemical molecule. In the last step of typical drug development procedure all potential drug targets have to be clinically validated to confirm the development of drugs for a particular protein target is worth proceeding.

We will provide shortly the specialized web pages that will present the human drug targets, i.e. 'druggable' subset of human proteome. Our resources will include human 'druggable' sequences, their crystal structures if known, 3D models for those sequences, sets of known inhibitors for each human drug target, the 3D structure (known or predicted) of protein-inhibitor complexes, together with detailed biological and chemical information on both human drug targets and their inhibitors. This database can be used as the core training set for our machine learning algorithms and global/local sequence similarity searches for further refinement of the set of potential drug targets. This resource will allow for rapid identification of most interesting 'druggable' targets in human proteome. According to Russ and Lampel [18] one decade from now the pharmaceutical industry challenge will lay in discovering the therapeutic effect of known leads and druggable targets. What is most important, in order to increase the size of our pharmacopeia and to cure human diseases more efficiently, new 'druggable' protein families must be identified.

Acknowledgements This work was supported by EC BioSapiens (LHSG-CT-2003–503265) and EC SEPSDA (SP22-CT-2004–003831) 6FP projects as well as the Polish Ministry of Education and Science (PBZ-MNiI-2/1/2005 and MNII ordinary research grant to DP).

References

1. Issac B, Raghava GP (2002) Biotechniques 33(3):548–550, 552, 554–6
2. Miller PL, Nadkarni PM, Carriero NM (1991) Comput Appl Biosci 7(1):71–78
3. Pearson WR (1990) Methods Enzymol 183:63–98
4. Pearson WR (1994) Methods Mol Biol 24:307–331
5. Altschul SF, Koonin EV (1998) Trends Biochem Sci 23(11):444–447
6. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25(17):3389–3402
7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) J Mol Biol 215(3):403–410
8. Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF (2006) BMC Biol 4:41
9. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Nucleic Acids Res 29 (14):2994–3005
10. Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) Bioinformatics 19(8):1015–1018
11. Kryshtafovych A, Venclovas C, Fidelis K, Moult J (2005) Proteins 61 Suppl 7:225–236
12. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A (2005) Proteins 7(61 Suppl):3–7
13. Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L (2004) Nucleic Acids Res 32(Web Server issue):W576–W581
14. Plewczynski D (2007) Adv Chem Info 1(1):11–19
15. Drews J, Ryser S (1997) Nat Biotechnol 15(13):1318–1319
16. Drews J (1996) Nat Biotechnol 14(11):1516–1518
17. Hopkins AL, Groom CR (2002) Nat Rev Drug Discov 1(9):727–730
18. Russ AP, Lampel S (2005) Drug Discov Today 10(23–24):1607–1610
19. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Graf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, Parker A, Proctor G, Prlic A, Rae M, Rios D, Redmond S, Schuster M, Sealy I, Searle S, Severin J, Slater G, Smedley D, Smith J, Stabenau A, Stalker J, Trevanion S, Ureta-

Vidal A, Vogel J, White S, Woodwark C, Hubbard TJ (2006) Nucleic Acids Res 34(Database issue):D556–D561
20. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E (2007) Nucleic Acids Res 35(Database issue):D610–D617
21. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2007) Nucleic Acids Res 35(Database issue):D224–D228
22. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, Wu CH (2005) Nucleic Acids Res 33(Database issue):D201–D205
23. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) Nucleic Acids Res 30(1):276–280
24. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL (2000) Nucleic Acids Res 28(1):263–266
25. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) Nucleic Acids Res 32 (Database issue):D138–D141
26. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Nucleic Acids Res 34(Database issue):D247–D251
27. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J (2006) Nucleic Acids Res 34 (Database issue):D668–D672
28. Li W, Godzik A (2006) Bioinformatics 22(13):1658–1659