Mutation tagging with gene identifiers applied to membrane protein stability prediction

Rainer Winnenburg, Conrad Plake, and Michael Schroeder
Biotec, TU Dresden, Germany
ms@biotec.tu-dresden.de

Abstract

The automated retrieval and integration of information about protein point mutations in combination with structure, domain and interaction data from literature and databases promises to be a valuable approach to study structure-function relationships in biomedical data sets.

As a prerequisite, we developed a rule- and regular expression-based protein point mutation retrieval pipeline for PubMed abstracts, which shows an F-measure of 87% for the pure mutation retrieval task on a benchmark dataset.

In order to link mutations to their proteins, we utilised a named entity recognition algorithm for the identification of gene names co-occurring in the abstract, and established links based on sequence checks. We identified more than 10Mio genes/proteins in nearly 3.5Mio abstracts and 260.000 mutations in 80.000 of these abtracts (2.3%). In 52% of cases the identified gene's sequence and the mutation are consistent. We evaluated the use of mutations in gene identification in detail on a small test set of 22 abstracts. Identifying the correct gene improved from 77% to 91% when considering the mutations.

To demonstrate practical relevance, we set up a mutation screening for five membrane proteins from the family of G protein-coupled receptors to evaluate a solvation energy based model for the prediction of stabilising regions in membrane proteins. We identified 35 mutations in text. 25 out of 35 mutation phenotypes reported in literature were in compliance with the prediction of the energy model, which supports a relation between mutations and stability issues in membrane proteins.

1 Introduction

Proteins carry out most cellular functions as they are acting as building blocks for structures, enzymes, gene regulators, and are involved in cell mobility and communication (Alberts et al., 2002). Proteins may interact briefly with each other in an enzymatic reaction, or for a long time to form part of a protein complex. The interactions between proteins are of central importance for almost all processes in living cells, and are described by numerous distinct pathways in databases such as KEGG (Ogata et al., 1999). Malfunctions or alterations in such pathways can be the cause of many diseases, when for instance the biosynthesis of involved proteins is repressed or proteins are not interacting the way they should. The latter can be due to structural changes in one of the interacting proteins, caused by point mutations, i.e. single wild type amino acid substitutions. Indeed, it is already well known that such mutations are the cause of many hereditary diseases. Thus the large-scale analysis of point mutation data in combination with information about protein interactions, protein structure and disease pathogenesis, might facilitate the study of still unresolved phenotypes and diseases.

It is envisaged to provide an automated system for the interpretation of structure-function relations in the context of genetic variability data. Despite the availability of numerous biomedical data collections, valuable information about mutation-phenotype associations is still hidden in non-structured text in the biomedical literature. Thus text mining methods are implemented to automatically retrieve these data from the 18 millions of literature references in PubMed. The extracted knowledge will be stored in one homogeneous data store and integrated with already available data from suitable databases. On the basis of all these combined data, new hypotheses can be formulated, like the prediction of phenotypic effects induced by mutations. At the moment, we are populating a database with organism specific protein-mutation associations which we envisage to apply on diverse biological problems, such as the detection of mutation centred gene-disease associations in human.

2 Background

Genomic variation data has already been collected for many years. Single nucleotide polymorphisms (SNPs), which make up about 90% of all human genetic variation and occur every 100 to 300 bases along the 3-billion-base human genome, are available as large collections. Single amino acid polymorphisms (SAPs) are often manually extracted from literature and curated into databases, originating from wet lab experiments. Additionally, some structures of such mutations may be revealed in crystallography experiments and might eventually end up as distinct structures in the Protein Database PDB. Of particular interest is the identification of mutations which have a strong influence on the stability of proteins. Therefore, the biomedical literature can be systematically searched for information about mutation-phenotype associations by text mining, which may lead to new insights beyond information in existing databases. For the text mined data it is additionally possible to weight or prioritise information according to their publication date, the involved authors and the journal. Considering these meta data can be relevant if for instance an already published assumption has been proven wrong in a more recent publication, or for determining whether a protein is a hot topic or if the information is already available for years. Furthermore, it is possible to receive a more detailed view on a protein's characteristics, e.g. if a certain interaction only takes place under specific conditions, or if an interaction is prevented by the conformational change of a protein domain triggered by a point mutation.

2.1 Databases

Data on mutations have been collected for years, for numerous species and by different organisations for diverse purposes. There are many efforts to cope with the data, which is being made available in a growing number of databases. The Human Genome Variation society (Horaitis and Cotton, 2004) promotes the collection, documentation and free distribution of genomic variation information. New mutation databases are reported in the Journal Human Mutation on a regular basis. There are manually curated databases like OMIM (Hamosh et al., 2002), UniProt Knowledgebase (Yip et al., 2008; Yip et al., 2007), and general central repositories like the Human Gene Mutation Database (Stenson et al., 2008), Universal Mutation Database (Broud et al., 2000), Human Genome Variation Database (Fredman et al., 2004), MutDB (Singh et al., 2007).

Besides these central repositories, there are small specialised databases, such as the infevers autoinflammatory mutation online registry (Milhavet et al., 2008), the GPCR NaVa database for natural variants in human G protein-coupled receptors (Kazius et al., 2007), or the Pompe disease mutation database with 107 sequence variants (Kroos et al., 2008).

In contrast, unpublished SNPs normally make their way into large locus specific data repositories. Since August 2006, there is a wiki based approach SNPedia in contrast to classical databases collecting information on variations in human DNA.

2.2 Text mining

Despite the availability of numerous biomedical data collections, valuable information about mutation-phenotype associations is still hidden in non-structured text in the biomedical literature. Thus text mining methods are implemented to automatically retrieve these data from the 18 millions of referenced articles in PubMed. Text mining aims to automatically extract and combine information spread

in several natural language texts and by this generating new hypotheses. One of the key prerequisites for finding new *facts* (e.g. *interactions* or *mutations*) is the named entity recognition (NER) in text, the assignment of a class to an entity (e.g. *protein*), as well as a preferred term or identifier, in case an entry in a database, such as *UniProt*, or a controlled vocabulary like the *Gene Ontology (GO)* (Ashburner et al., 2000) exists. For the task of named entity recognition usually a dictionary is used, which contains a list of all known entity names of a class (e.g. human proteins) including synonyms. For the recognition of patterns (e.g. database identifiers like *NM_12345*) regular expression can be defined. For the analysis of whole sentences, *Natural language processing (NLP)* techniques are used, which aim to understand text on a syntactic and semantic level. This approach is often paired with systems which are based on a set of manually defined *rules* or which make use of (semi-)supervised *machine learning* algorithms.

Up to now, there have already been diverse examples for the successful application of text mining to the mutation retrieval task. Early examples are the automatic extraction of mutations from Medline and cross-validation with OMIM (Rebholz-Schuhmann et al., 2004), and the work by (Cantor and Lussier, 2004), who mined OMIM for phenotypic and genetic information to gain insights into complex diseases. More recently, (Caporaso et al., 2007b) applied their concept recognition system based on regular expressions on mutation mining task, and the automatic Extraction of Protein Point Mutations Using a Graph Bigram association (Lee et al., 2007) was reported to find reliably gene-mutation associations in full text. For identifying gene-specific variations in biomedical text, (Klinger et al., 2007) integrate the ProMiner system developed for the recognition and normalisation of gene and protein names with a conditional random field (CRF)-based recognition system. As an answer to the diverse approaches developed over the past years, a framework for the systematic analysis of mutation extraction systems was proposed by (Witte and Baker, 2007).

More and more groups are working on mutations in proteins and their involvement in diseases. (Kanagasabai et al., 2007) developed mSTRAP (Mutation extraction and STRucture Annotation Pipeline), for mining mutation annotations from full-text biomedical literature, which they subsequently used for protein structure annotation and visualisation. (Worth et al., 2007) use structure prediction to analyse the effects of nonsynonymous single nucleotide polymorphisms (nsSNPs) with regard to diseases. Focussing on Alzheimer's disease, (Erdogmus and Sezerman, 2007) extract mutation-gene pairs, with estimated 91.3%, and precision at 88.9%. (Lage et al., 2007) realised a human phenome-interactome network of protein complexes implicated in genetic disorders by by integrating quality-controlled interactions of human proteins with a validated, computationally derived phenotype similarity score,

3 Methods

Through the combination of different data from literature and databases it is possible to derive new facts, e.g. novel gene-disease associations or the influence of mutations on protein-protein interactions. The approach is designed in such a way, that it can in principle be applied to any kind of genetic data for answering disease centred questions. For the moment, we concentrate on collecting available high quality data on protein point mutations from curated databases and from peer-reviewed literature. For the latter we will present a flexible approach for both the specific and high-throughput retrieval of mutations. In detail, the following tasks have to be performed: (1) Identify genes/ proteins in abstracts. (2) From this subset consider only these which additionally contain information about mutations. (3) Propose potential protein - mutation pairs. (4) Filter proposed pairs by sequence compliance. (5) Utilise this information for the refinement of the original gene/protein identifier.

3.1 Entity recognition

Gene normalisation This module allows for the automated named entity recognition of genes and proteins. Our approach performs gene name disambiguation by using background knowledge to match a gene with its context against the text as a whole (Hakenberg et al., 2007). A gene's context contains information on Gene Ontology annotations, functions, tissues, diseases etc. extracted from the databases Entrez Gene and UniProt. A comparison

of gene contexts against the text gives a ranking of candidate identifiers and the top ranked identifier is taken if it scores above a defined threshold. This approach has been recently extended for inter-species normalisation and achieves 81% success rate on a mixed dataset of 13 species (Hakenberg et al., 2008).

Mutation tagging We implemented an entity recognition algorithm (*MutationTagger*) to automatically extract protein point mutation mentions from PubMed abstracts. Wild-type and mutant amino acid, as well as the sequence position of the substitution are extracted by means of both a set of regular expressions for pattern recognition of 1 or 3-letter-notations (e.g. *E312A* or *Glu(312)→Ala*), and rules for the more complex identification of textual mutation descriptions (e.g. *Glu312 was replaced with alanine*). Problems concerning the full text representations (detecting the correct sequence position of the mutated residue and unravelling enumerations) have been addressed by additional extraction algorithms and the implementation of a sequence check. An evaluation of our method on the test data from MutationFinder (Caporaso et al., 2007a) showed comparable success rates of around 89% F-measure for mutation mention extraction.

3.2 Association of entity pairs

In the process of recognising mutations in text, the normalisation, i.e. the direct association to specific proteins, remains a challenge. This is due to the fact that the abstracts of relevant publications typically mention more than only one single mutation and protein. Thus, a mutation-protein association purely based on their co-occurrence in one abstract is not sufficient, as it would result in a permutation with a huge number of false positive predictions. The problem becomes even more evident, when considering that both gene and mutation tagging are imperfect, achieving a precision of 80 to 90% each.

A method is desired, that both disambiguates the relations of candidate mutations and proteins, and filters out false positives from the underlying individual mutation and protein recognition tasks. There are approaches which apply a word distance metric for assigning a mutation to its nearest occurring protein term, which is error prone, as matching mutation and protein do not necessarily have to occur close to each other in the abstract or even in the same sentence. The statistical approach GraB is an excellent tool for the automatic extraction of Protein Point Mutations using a Graph Bigram association (Lee et al., 2007), achieving good results for most likely mutation-protein association but alone would also not fulfil the second aspect of filtering out false positives.

Sequence Checks Mutations are commonly described as the substitution of a wild-type by a mutant amino acid at a given position. Our method compares the wild-type residue as described in a mutation mention with the UniProt/Swiss-Prot and PDB protein sequences for all candidate proteins. It is important to incorporate sequences from both repositories, as the sequence numbering can differ and it is not always evident from a publication's abstract, which numbering the mutation notation refers to. To map UniProt IDs to PDB and vice versa, we used PDB cross-references in UniProtKB/Swiss-Prot from http://beta.uniprot.org/docs/pdbtosp and the residue specific comparison between PDB and SwissProt sequences as provided by http://www.bioinf.org.uk/pdbsws/ (Martin, 2005). Only associations between mutations and proteins with matching sequences are considered.

3.3 Annotation pipelines

The developed mutation retrieval pipeline can be accessed through two different interfaces (see Figure 1), which offer dependent on the annotation task, either a systematic or quick and flexible solution. The following approaches have been implemented:

- **Organism-centred approach (database)**

  All available mutations for a given organism will be retrieved in one single literature screening and stored in the Mutation database. This approach relies on the large-scale identification of gene mentions in PubMed abstracts, which have to be compiled for organisms of interest prior to a mutation screening. As of now, gene mention data is available for human, mouse, and yeast. However, data for additional relevant organisms will be added on a regular basis in the near future.

- **Protein-centred approach (on-the-fly)**

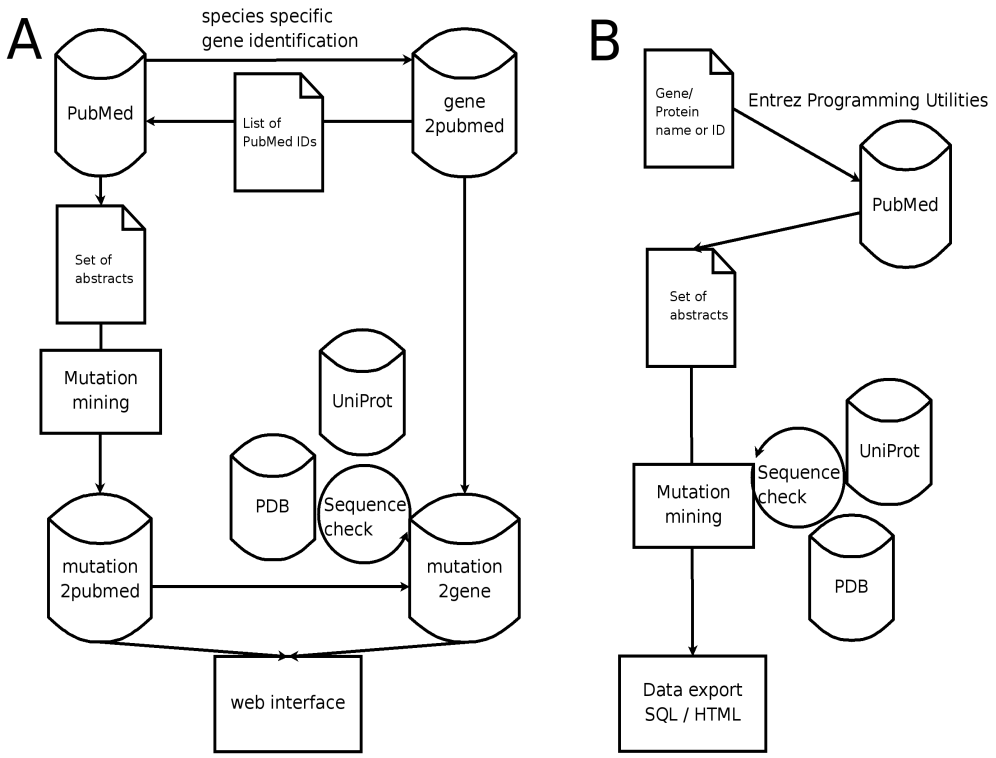

Figure 1: Workflow of mutation data retrieval with MutationTagger. A: abstracts mentioning proteins for given species are tagged for mutations. The filtered data is written to database. B: For a protein of interest relevant articles are retrieved and tagged for mutations. The filtered data can be exported to HTML or SQL.

It is possible to retrieve relevant data for a single gene or a list of genes/ proteins for any organism. For this purpose, the gene identification part performed by the gene normaliser is replaced by a direct full text search in the PubMed library using the Entrez Programming Utilities. Again, the result is a set of abstracts, which is subsequently processed by the MutationTagger.

3.4 Improvement of gene normalisation

As described above, we defined the input set of documents for the organism-centred mutation mining approach by scanning the whole PubMed database for abstracts mentioning at least one gene or protein of a pre-defined species. For this filtering step, we relied on the gene normalisation techniques of our gene normaliser, which was applied to all PubMed abstracts in advance and has shown 85% F-measure for human genes and slightly lower for other species. However, the gene normalisation proposes by default only one single identifier per gene mention, even if a set of different candidate identifiers was computed. According to internal ranking mechanisms, only the top scoring candidate is considered. This leads to a possible scenario, where in some cases the correct identifier is ranked lower and would be neglected for any subsequent data procession. In case of our mutation mining algorithm, we assume that some mutations cannot be associated to the correct protein, because the gene tagging task already failed.

On the other hand, it should be possible to improve the performance of both entity recognition techniques for genes and mutations by combining the results. The idea is to run both approaches with low precision thus receiving a high recall, permutate all elements of both sets, and then consider the intersection of all combinations that fit. Mutation and gene product are considered to be a valid pair, if the wild-type residues at the mutated position in the protein sequence and in the reported mutation match (as described in section 3.1). For all

proposed gene identifiers, protein sequences are obtained and checked for compliance with the reported wild type amino acid. The score of identifiers that show a match are increased, which might lead to a re-ranking of the identifiers for one gene entity. This could further improve the original gene normalisation approach for candidate entities which are reported to show a mutation.

Example As shown in Figure 2 our gene normaliser identified CCP (human crystallin, gamma D; EntrezGene ID 1421) as the top candidate gene name for abstract PMID 8142383. The mutation tagger identified a replacement of tryptophan with glycine at position 191 as the only mutation mentioned in the paper. None of the protein sequences retrieved for human CCP showed a tryptophan residue at position 191, which means that this gene identifier was not supported by mutation information. However, besides human crystallin, there was also cytochrome-c peroxidase in yeast (EntrezGene ID 853940) proposed as an alternative identifier, which received a lower score. As the product of this gene showed a tryptophan residue at postion 191 (according to PDB sequencing) the score was increased making it the new top candidate. Indeed, manual curation of the corresponding literature confirmed, that the only gene mentioned in the abstract is cytochrome-c peroxidase in yeast. The same positive re-ranking finding the correct gene identifier through mutation information was shown for human TP53 in paper 11254385, and human amylase alpha in paper 15182367.

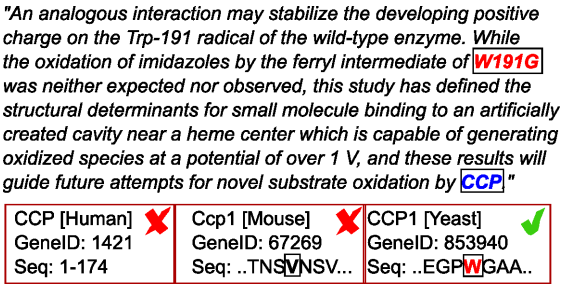

Figure 2: Example for gene name normalisation with the help of mutation mining. Initially, our gene normaliser proposed the human gene CCP as its context fits the text best (abstract not fully shown). However, when comparing the recognised mutation at position 191 with the sequences of all three candidates, only CCP in yeast contains the wild-type tryptophan at the specified position (PDB entry). After checking the full text of this publication, we found that CCP indeed refers to the gene in Saccharomyces cerevisiae.

4 Results

Mutation database In order to establish a mutation database, which will eventually store all protein point mutations mentioned in PubMed abstracts for all organisms of interest, a first platform has been realised, comprising a MySQL database, which can be accessed by a web-interface.

To populate the database, in a first step the PubMed corpus is filtered for abstracts mentioning at least one gene or protein using the named entity recognition algorithm as described in Section 3.1, which is currently working for the three organisms human, mouse, and yeast. This led to a set of set of 3,443,566 abstracts proposing more than 10 millions of potential protein candidates. In a second step, the mutation extraction algorithm is applied on this corpus and the retrieved information is transferred into the database. In total, 258,511 mutations were found in 78,968 abstracts. Subsequently, for all candidate genes found in these abstracts, the corresponding sequences are obtained and checked for compliance with the wild type amino acid at the position of the mentioned mutation, which led to a number of 877,183 potential protein - mutation pairs. Out of these, 127,384 are supported by sequence (74,722 if multiple mentions of the same mutation in one abstract are counted as one) in contrast to 131,127 (77,643) mutations which have not passed the sequence filter. In summary, from all mutations identified by the plain algorithm, about 49% could be supported by gene associations based on sequence check. These data were retrieved from 41,384 (52%) abstracts in total.

Evaluation We evaluated our approach on two different tasks: pure identification of a mutation in a text, and the identification of correct mutation-protein pairs. An evaluation of our method on the test data from MutationFinder (Caporaso et al., 2007a) showed comparable success rates of around 87% F-measure for pure mutation mention extraction. On the document level, from 182 abstracts con-

taining mutations, 163 were identified, in 4 abstracts mutation were wrongly predicted. On the mutation level 741 out of 907 were identified alongside 61 false positives.

To assess the refinement possibilities for falsely top ranked gene names, from the 182 abstracts we took the subset of those, the gene normaliser identified genes from one of the 10 supported species: human, mouse, yeast, rat, fruit fly, H. pylori, S. Pombe, C. Elegans, A. Thaliana, and D. Rerio. This led to a subset of 22 abstracts. In the initial run, the gene name identifier identified in 17 of 22 abstracts (77%) the correct gene as the top ranked candidate. However, after the gene tagging refinement by applying the sequence filter to all candidate genes, the genes of 3 more papers were identified correctly replacing the original and false top candidate. This led to the correct protein normalisation for 20 out of 22 (91%) publications. For the remaining 2 publication, the correct genes could not be identified, as they were from species, the gene identifier does not yet support. The suggested genes from mouse were first falsely predicted, which were then not supported by the sequence checks. By this the proposed identifiers were brought below the threshold, resulting in no gene identification at all for these 2 abstracts and turning the 2 "false positives" to "false negatives".

On-the-fly vs. database approach We evaluated the results of the two access approaches (database and on-the-fly) for human Aquaporin-1, as part of the stability analysis of protein membranes (see Section 5). The precision of the on-the-fly approach is expected to be lower, as the first step is more general due to relying on full text searches instead of entity recognition. Indeed, in comparison to the unique 20 mutations found by the organism-centred approach, 9 additional mutations were found, of which all were false positives, actually appearing in Aquaporin-2 or 4. This supports the good precision of the named entity approach for the gene normalisation.

5 Application

Predicting effects of mutations based on sequence Integral membrane proteins play an important role in all organisms, especially as transporters. Due to their striking importance, mutations in membrane proteins are known to be the cause of many hereditary diseases, such as cystic fibrosis, or retinitis pigmentosa. The reason are often conformational changes in proteins, which may lead to malfunction of a whole protein complex. Unfortunately, identified structures for membrane proteins are still rare. For this reason, we used a coarse grained model presented by (Dressel et al., 2008) considering sequence information only, to assess the influence of mutations on protein structure.

The approach considers the solvation energy, which is based on the probability distribution for each amino acid within the integral part of a membrane protein to be facing the membrane or other proteins. The amino acid specific property inside or outside reflects the orientation of the amino acid side chains with respect to the centre of mass of the neighbouring residues. For a given mutation, the approach compares the solvation energies for wild-type and mutant residues. If the energies differ significantly, a destabilising effect is predicted, especially if the energies are changing from negative to positive or vice versa.

To quantify the ability of this model to predict the influence of mutations on the stability of membrane proteins, we compared already examined and published effects of mutations with the predictions of the sequence based model. For this purpose, we screened the literature for single point mutations reported for five membrane proteins from the family of G protein-coupled receptors (bacteriorhodopsin and halorhodopsin from *Halobacterium salinarum*, bovine rhodopsin, Na+/H+ antiporter from *Escherichia coli*, and human aquaporin-1). As described in Section 4, *Protein-centred approach* and Figure 1B, articles relevant for these proteins were identified by searching PubMed via the NCBI Entrez Programming Utilities. Abstracts for each protein were queried by the protein and gene name including the synonyms as derived from the corresponding PDB/UniProt entry.

The MutationTagger was applied on these five sets of abstracts for the extraction of mutation information. The application of sequence checks brought the results down to a reasonable number of proposed mutations, which were presented as HTML documents and subsequently manually curated. We only used the publications where a single point mutation was discussed in the context of stability or stabil-

ity related function. Double or multiple mutations were not considered, as the determination of a direct relation between the reported effect and one of the mutations is not possible. If an appropriate mutation was found in the literature, we compared the solvation energies of both wild-type and mutant residues to decide, if the mutation was stabilising, slightly stabilising, slightly destabilising, or destabilising.

Example Mutation T93P for bovine rhodopsin was reported to lead to a conformational change of the protein. Considering the two solvation energies of wild type Threonine (-0.66 a.u.) and mutant Proline (0.08 a.u.) a destabilising effect can be predicted, although both amino acids are actually classified as neutral. Without the change of sign from - to +, an only slightly destabilising effect would have been hypothesised.

Relevance We were able to show the ability of our mutation mining approach to retrieve publications containing mutation information for given proteins at a good precision. Due to the quick and precise retrieval of mutation data we were able to assess the soundness of the coarse grained model for the prediction of stabilising regions in membrane proteins. 25 out of 35 mutational effects reported in the literature for any of these five membrane proteins correlate with the predictions based on the solvation energy. These cases suggest a relation between mutations and stability issues in membrane proteins.

Acknowledgement: We are grateful for financial support by the EU project Sealife and the BMBF Format Project CLSD and to Frank Dressel and Dirk Labudde for discussions on the application.

6 Conclusion

We developed a rule- and regular expression-based approach that allows for the retrieval of protein point mutations from the whole PubMed database specifically for any given protein. This flexibility makes it a powerful tool for immediately finding relevant data for follow-up studies, as we showed in the application on five membrane proteins. In addition, MutationTagger can be utilised for the species-wide identification of mutations in proteins mentioned in PubMed. We started to set up a mutation database which allows for systematically querying mutation related information, and finding relevant literature for subsequent studies. The sequence checks applied on identified mutations and candidate proteins have been proven to be an efficient, yet not sufficient filter for determing mutation-protein associations. The filter shows good sensitivity but improvable specificity, especially regarding the species level. Furthermore, we were able to show, that the mutation information from literature can even further improve the quality of the gene tagging algorithm we used, which already showed very good results.

References

B Alberts, D Bray, K Hopkin, A Johnson, J Lewis, M Raff, K Roberts, and P Walter. 2002. *Essential Cell Biology*. Garland Science Textbooks, London.

Michael Ashburner, Catherine Ball, Judith Blake, David Botstein, Heather Butler, J. Cherry, Allan Davis, Kara Dolinski, Selina Dwight, Janan Eppig, Midori Harris, David Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John Matese, Joel Richardson, Martin Ringwald, Gerald Rubin, and Gavin Sherlock. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics.*, 25:25–29, May. 10.1038/75556.

C Broud, G Collod-Broud, C Boileau, T Soussi, and C Junien. 2000. Umd (universal mutation database): a generic software to build and analyze locus-specific databases. *Hum Mutat*, 15(1):86–94.

MN Cantor and YA Lussier. 2004. Mining omim for insight into complex diseases. *Medinfo*, 11(Pt 2):753–7.

J. Gregory Caporaso, Jr William A. Baumgartner, David A. Randolph, K. Bretonnel Cohen, and Lawrence Hunter. 2007a. Mutationfinder: A high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23:1862–1865, Jul. 10.1093/bioinformatics/btm235.

J. Gregory Caporaso, William A. Baumgartner, David A. Randolph, K. Bretonnel Cohen, and Lawrence Hunter. 2007b. Rapid pattern development for concept recognition systems: application to point mutations. *Journal of bioinformatics and computational biology*, 5:1233–1259, Dec.

Andreas Doms and Michael Schroeder. 2005. Gopubmed: exploring pubmed with the gene ontology. *Nucleic Acids Res*, 33:W783–6, Jul. 10.1093/nar/gki470.

F Dressel, A Marsico, A Tuukkanen, R Winnenburg, D Labudde, and M Schroeder. 2008. Stabilizing regions in membrane proteins. In *From Computational*

Biophysics to Systems Biology (CBSB08), pages 197–9.

M Erdogmus and OU Sezerman. 2007. Application of automatic mutation-gene pair extraction to diseases. *J Bioinform Comput Biol*, 5(6):1261–75, Dec.

D Fredman, G Munns, D Rios, F Sjholm, M Siegfried, B Lenhard, H Lehvslaiho, and AJ Brookes. 2004. Hgvbase: a curated resource describing human dna variation and phenotype relationships. *Nucleic Acids Res*, 32(Database issue):D516–9, Jan.

Jörg Hakenberg, Loic Royer, Conrad Plake, Hendrik Strobelt, and Michael Schroeder. 2007. Me and my friends: gene mention normalization with background knowledge. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 141–4.

J Hakenberg, C Plake, R Leaman, M Schroeder, and G Gonzales. 2008. Inter-species normalization of gene mentions with GNAT. *Bioinformatics*. to appear.

A Hamosh, AF Scott, J Amberger, C Bocchini, D Valle, and VA McKusick. 2002. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 30(1):52–5, Jan.

O Horaitis and RG Cotton. 2004. The challenge of documenting mutation across the genome: the human genome variation society approach. *Hum Mutat*, 23(5):447–52, May.

R Kanagasabai, KH Choo, S Ranganathan, and CJ Baker. 2007. A workflow for mutation extraction and structure annotation. *J Bioinform Comput Biol*, 5(6):1319–37, Dec.

J Kazius, K Wurdinger, Iterson M van, J Kok, T Bck, and AP Ijzerman. 2007. Gpcr nava database: natural variants in human g protein-coupled receptors. *Hum Mutat*, Oct.

R Klinger, CM Friedrich, HT Mevissen, J Fluck, M Hofmann-Apitius, LI Furlong, and F Sanz. 2007. Identifying gene-specific variations in biomedical text. *J Bioinform Comput Biol*, 5(6):1277–96, Dec.

M Kroos, RJ Pomponio, Vliet L van, RE Palmer, M Phipps, der Helm R Van, D Halley, and A Reuser and. 2008. Update of the pompe disease mutation database with 107 sequence variants and a format for severity rating. *Hum Mutat*, Apr.

K Lage, EO Karlberg, ZM Strling, PI Olason, AG Pedersen, O Rigina, AM Hinsby, Z Tmer, F Pociot, N Tommerup, Y Moreau, and S Brunak. 2007. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–16, Mar.

Lawrence C. Lee, Florence Horn, and Fred E. Cohen. 2007. Automatic extraction of protein point mutations using a graph bigram association. *PLoS computational biology*, 3:e16, Feb. 10.1371/journal.pcbi.0030016.

AC Martin. 2005. Mapping pdb chains to uniprotkb entries. *Bioinformatics*, 21(23):4297–301, Dec.

F Milhavet, L Cuisset, HM Hoffman, R Slim, H El-Shanti, I Aksentijevich, S Lesage, H Waterham, C Wise, de Menthiere C Sarrauste, and I Touitou. 2008. The infevers autoinflammatory mutation online registry: update with new genes and functions. *Hum Mutat*, Apr.

H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. 1999. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27:29–34, Jan.

D Rebholz-Schuhmann, S Marcel, S Albert, R Tolle, G Casari, and H Kirsch. 2004. Automatic extraction of mutations from medline and cross-validation with omim. *Nucleic Acids Res*, 32(1):135–42.

A Singh, A Olowoyeye, PH Baenziger, J Dantzer, MG Kann, P Radivojac, R Heiland, and SD Mooney. 2007. Mutdb: update on development of tools for the biochemical analysis of genetic variation. *Nucleic Acids Res*, Sep.

PD Stenson, E Ball, K Howells, A Phillips, M Mort, and DN Cooper. 2008. Human gene mutation database: towards a comprehensive central mutation database. *J Med Genet*, 45(2):124–6, Feb.

R Witte and CJ Baker. 2007. Towards a systematic evaluation of protein mutation extraction systems. *J Bioinform Comput Biol*, 5(6):1339–59, Dec.

CL Worth, GR Bickerton, A Schreyer, JR Forman, TM Cheng, S Lee, S Gong, DF Burke, and TL Blundell. 2007. A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nssnps) and their relation to disease. *J Bioinform Comput Biol*, 5(6):1297–318, Dec.

YL Yip, N Lachenal, V Pillet, and AL Veuthey. 2007. Retrieving mutation-specific information for human proteins in uniprot/swiss-prot knowledgebase. *J Bioinform Comput Biol*, 5(6):1215–31, Dec.

YL Yip, M Famiglietti, A Gos, PD Duek, FP David, A Gateau, and A Bairoch. 2008. Annotating single amino acid polymorphisms in the uniprot/swiss-prot knowledgebase. *Hum Mutat*, Jan.