BMC Bioinformatics

Introduction

Between proteins and phenotypes: annotation and interpretation of mutations

Christopher JO Baker*[1] and Dietrich Rebholz-Schuhmann[2]

Address: [1]University of New Brunswick, Saint John, PO Box 5050, Canada, E2L 4L5 and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK

Email: Christopher JO Baker* - bakerc@unb.ca; Dietrich Rebholz-Schuhmann - rebholz@ebi.ac.uk

* Corresponding author

from ECCB 2008 Workshop: Annotations, interpretation and management of mutations (AIMM)
Cagliari, Italy. 22 September 2008

Published: 27 August 2009

BMC Bioinformatics 2009, **10**(Suppl 8):I1 doi:10.1186/1471-2105-10-S8-I1

This article is available from: http://www.biomedcentral.com/1471-2105/10/S8/I1

Understanding the roles of genes and proteins and the functional consequences of their mutation is mandatory for the interpretation of a resulting phenotype, for example the observed disease state. Biomedical scientists are approaching this task from different angles. Typically individual research projects investigate the consequences of natural and experimental mutations in proteins by analysing stability and the resulting changes to a protein's function [1]. The consequences of such changes may be used in simulations and subsequent population studies may analyse cohorts for genetic variability, after which review of the observed phenotypes in each of the cohort's individuals may occur [2,3]. Results are reported in the scientific literature and in biomedical databases but are not necessarily integrated according to a detailed and comprehensive knowledge representation, nor are they accessible for reuse by multiple stakeholders.

Annotation of mutations with their relevance for phenotypic expression is crucial to the understanding of genetic mechanisms, biological processes and complex diseases. Large-scale resources able to cope with the full extent of data and annotations reporting on human variability (see 1000 genomes project) have yet to be built [4]. Systems maintaining this data will have to deal with a number of problems: (i) extraction, storage and reuse of genotype-phenotype information [5,6], (ii) preparation of semantic resources for phenotype description (phenotype ontologies, e.g. human phenotype ontology) [7,23], and (iii) automated interpretation, simulation and prediction of functional changes induced by mutations and sequence variants [8].

In principle, a complete solution that allows efficient hypothesis generation based on annotated mutation information would be able to derive, from the location of the mutation, changes in protein structure and function thus allowing prediction of changes in the protein's activity and the molecular processes that are affected. Finally, it would realize and describe the overall changes in the microscopic, physiological and macroscopic phenotype. Such a solution is not yet in sight.

The integration of data from different experimental and simulation methods as well as the annotation of data with rich semantic resources (bio-ontologies [9]) are the most important next steps [10]. To date, moving from SNP to sequence to structure and function has been addressed with varying degrees of accuracy with sequence and structure based methods. The need to apply prediction techniques at a genomic scale requires that adequate solutions have to be identified and benchmarked against reliable measures. Only then can we anticipate and design solutions to address changes expected from significant mutations. In this context the reuse of existing mutations and annotations, from databases and those mined from the literature, for checking the quality of predictions is pivotal.

In order to assess the state of the art in the annotation, interpretation, management of mutations several initiatives and events have recently taken place [11,12]

These events have brought to light the challenges, the existing solutions and relevant expertise within the research community. This supplement to *BMC Bioinformatics* on mutation related IT solutions gives further insight on the current state of research in this important domain. The presented research is mainly concerned with the impact of mutations on protein stability and protein function but also includes reports on attempts to predict the relevance of a protein's modifications in the context of the disease under investigation. Several publications report on the annotation of mutations in the corresponding gene/protein with information extracted from the scientific literature. A number of the studies used specific protein families to demonstrate the usefulness of the findings.

It is clearly an ambitious goal to discover novel knowledge from the literature by contrasting the retrieved content against reference data resources [13]. For the protein kinases, [14] processed a large set of documents to identify mutations and validate the results against against KinMutBase, revealing an overlap of only 52%. In other approaches, contextual information in neighbouring text was used to find functional annotations of protein residues that could be evaluated against UniProtKb [15] or to efficiently retrieve documents reporting on mutations that destabilise G protein-coupled receptors [16]. In the latter case, the reported destabilisation effects were evaluated against sequence based predictors. Contextual information in Medline abstracts was also exploited, using different machine learning based classifiers, to predict the stability and the disease relevance of mutations in lipase and amylase enzymes [17].

Two further publications describe the integration of extracted mutations into solutions that support interpretation of SNPs in a larger context. Mutations can be visualized in a 3D representation of the protein's structure to better understand the induced effects of the mutation through its three-dimensional localisation [18]. In another approach, the impact of changes to a protein's activity due to its mutation can be traced through protein network simulation [2]. The authors use the example of MEK1 activation of Akt to demonstrate the usefulness of this approach. Both publications show the complexity of the integration tasks, reveal new methodologies – based in part on existing infrastructure, and illustrate new biological insights in the context of existing biological models.

In parallel, other researchers seek to predict the consequence of sequence variation on biological processes. One research team has focused on mutations in mitochondrial sequences and their relevance for diseases. They propose and evaluate a scoring function that has been tested against other prediction methods (SIFT, PolyPhen, PHD-SNP, PLHOST) [19]. In another study, changes in a protein sequence and the protein's stability were assessed in the context of changes in the protein's function – which is assumed to be crucial to the development of a disease state [1]. These researchers showed that changes to the protein's stability have an effect on the protein's function, but there are also cases where the function is not altered at all. In a similar approach, another team used features from protein annotations (e.g., from UniProt) to classify mutations so as to predict changes in the protein's function and its relevance for diseases [20]. Clearly, taken in isolation the criterion of large scale change in a protein's stability is not sufficient to distinguish deleterious mutations and neutral variation. Finally, [21] analyse the adaptations of the HIV virus during the drug treatment of the patients. The genetic variability in the virus' protease is evaluated under the evolutionary pressure induced by the treatment of the patients with a varying number of antiviral drugs.

Altogether, this special issue gives an overview on the scope of ongoing research to exploit resources reporting on genetic variability. The full integration of all such resources is work in progress and clearly a necessary research direction with significant impact on many areas of biomedical science. Through a renewed community focus on life science infrastructure [9,22] there are opportunities to define a roadmap for technology development in this domain and to evolve existing approaches into a robust framework for phenotype prediction, based on annotations and interpretation of mutation data. Clearly many stakeholders and skills sets are required and a strong community focus must be present.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 8, 2009: Proceedings of the European Conference on Computational Biology (ECCB) 2008 Workshop: Annotation, interpretation and management of mutations. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/10?issue=S8.

References

1. Bromberg Yana, Rost Burkhard: **Correlating protein function and stability through the analysis of single amino acid substitutions.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S8.
2. Bauer-Mehren Anna, Furlong Laura, Rautschka Michael, Sanz Ferran: **From SNPs to pathways: integration of functional effect of sequence variations on models of cell signalling pathways.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S6.
3. Becker K: **The Gene Association Database.** *Nature Genetics* 2004, **36(5):**431-432.
4. **Cite 1000 Genomes** [http://www.1000genomes.org/]
5. Baker CJO, Witte R: **Mutation Mining-A Prospector's Tale.** *Information Systems Frontiers* 2006, **8:**47-57.
6. Kanagasabai R, Choo KH, Ranganathan S, Baker CJO: **A Work flow for Mutation Extraction and Structure Annotation.** *J Bioinformatics and Computational Biology* 2007, **5(6):**1319-1337.
7. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S: **The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease.** *Am J Hum Genet* 2008, **83(5):**610-5.
8. Bromberg Y, Rost B: **SNAP: predict effect of non-synonymous polymorphisms on function.** *Nucleic Acids Res* 2007, **35(11):**3823-3835.
9. **National Centre for Biomedical Ontology** [http://bioontology.org]
10. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nature Biotech* 2007, **25(11):**1251-5.
11. [http://www.ebi.ac.uk/Rebholz-srv/aimm.html].
12. Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, Kirsch H: **Automatic extraction of mutations from Medline and cross-validation with OMIM.** *Nucleic Acids Res* 2004, **32(1):**135-42.
13. Baker CJO: **Introduction. Making Sense of Mutations Requires Knowledge Management.** *J Bioinformatics and Computational Biology* 2007, **5(6):**v-vii.
14. Krallinger Martin, Izarzugaza Jose MG, Rodriguez-Penagos Carlos, Valencia Alfonso: **Extraction of human kinase mutations from literature, databases and genotyping studies.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S1.
15. Nagel Kevin, Yepes Antonio Jimeno, Rebholz-Schuhmann Dietrich: **Annotation of protein residues based on a literature analysis: cross-validation against UniProtKb.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S4.
16. Winnenburg Rainer, Plake Conrad, Schroeder Michael: **Improved mutation tagging with gene identifiers applied to membrane protein stability prediction.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S3.
17. Yeniterzi Süveyda, Sezerman Uğur: **EnzyMiner: automatic identification of protein level mutations and their impact on target enzymes from PubMed abstracts.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S2.
18. Izarzugaza Jose MG, Baresic Anja, McMillan Lisa EM, Yeats Corin, Clegg Andrew B, Orengo Christine A, Martin Andrew CR, Valencia Alfonso: **An integrated approach to the interpretation of Single Amino Acid Polymorphisms within the framework of CATH and Gene3D.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S5.
19. Bhardwaj Anshu, Mukerji Mitali, Sharma Shipra, Paul Jinny, Gokhale Chaitanya S, Srivastava Achal K, Tiwari Shrish: **MtSNPscore: a combined evidence approach for assessing cumulative impact of mitochondrial variations in disease.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S7.
20. Reumers Joke, Schymkowitz Joost, Rousseau Fréderic: **Using structural bioinformatics to investigate the impact of non synonymous SNPs and disease mutations: scope and limitations.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S9.
21. Haq Omar, Levy Ronald M, Morozov Alexandre V, Andrec Michael: **Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease.** *BMC Bioinformatics* 2009, **10(Suppl 8):**S10.
22. **Elixir** [http://www.elixir-europe.org/]
23. Sam LT, Mendona EC, Li J, Blake J, Friedman C, Lussier YA, Phen GO: **An integrated resource for the multiscale mining of clinical and biological date.** *BMC Bioinformatics* 2009, **10(Suppl 2):**S8.