

Mutation tagging with gene identifiers applied to membrane protein stability prediction

Rainer Winnenburg, Conrad Plake, and Michael Schroeder

AIMM workshop @ ECCB'08

September 22, 2008

The aim of our studies

- Question 1:
 - How can gene and mutation identification improve each other?
- Question 2:
 - How can mutation mining help to validate predicted energy barriers in membrane proteins?

Mutations and diseases

- single nucleotide polymorphisms (SNPs)
- amino acid replacement (SAPs)
- change of a molecule conformation
- disruption of proper interdomain interactions
- alterations in pathways
- malfunction can result in diseases
- treatment strategies

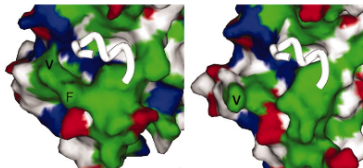


Normal red blood cell



Sickled red blood cell

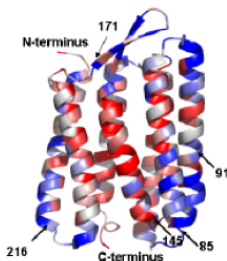
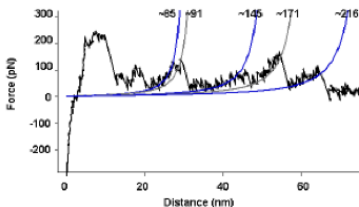
Local structure at the site of Phe-508 in NBD1 of CFTR



Lewis, H. A. et al. J. Biol. Chem. 2005;280:1346-1353

Prediction of stabilising regions in membrane proteins

- single molecule force spectroscopy experiments: pull protein out of membrane and measure force
- introduced mutations to study destabilising effects
- model that predicts which residues are potential targets
- trained and tested on known mutations in membrane proteins



Question 1: Cycle of gene and mutation identification

on a paper-by-paper basis:

- 1 Gene identification
find and identify genes/ proteins in abstracts
- 2 Mutation identification
find mutations in abstracts
- 3 Propose potential protein - mutation pairs
filter by protein sequence checks
- 4 Improvement of gene identification
mutation information refines original gene/protein identifiers

Gene identification

- gene name disambiguation using background knowledge
- ranking of candidate identifiers
- 81% success rate on a mixed dataset of 13 species

A gene encoding a putative human RNA helicase, p54, has been cloned and mapped to the band q23.3 of chromosome 11. The predicted amino acid sequence shares a striking homology (75% identical) with the female germline-specific RNA helicase ME31B gene of *Drosophila*. Unlike ME31B, however, the new gene expresses an abundant transcript in a large number of adult tissues and its 5' non-coding region was found split in a t(11;14) (q23.3;q32.3) cell line from a diffuse large B-cell lymphoma.

P54; RCK; HLR2

Species: *H. sapiens*

Chromosome 11q23.3

GO: RNA helicase

P54; NMT55; NRB54

Species: *H. sapiens*

Chromosome Xq13.1

GO: RNA splicing

P54; FKBP51; PPlase

Species: *H. sapiens*

Chromosome 6p21.3-2

GO: isomerase activity

S4; dRpt2; p54;p56

Species: *D. melanogaster*

Chromosome 3R:95C13

GO: Proteolysis

Hakenberg et al. *Genome Biology* 2008 9(Suppl 2):S14

@ECCB'08: Wednesday, 12:10pm

Mutation identification

.. potential sites for O-glycosylation. **Mutagenesis of asparagine 22 to glutamine abolished N-linked glycosylation of the V2 receptor (N22Q-V2R), without altering its function or level of expression.** The N22Q-V2R expressed in transfected cell migrated in denaturing ..

Mutation identification

.. potential sites for O-glycosylation. **Mutagenesis of asparagine 22 to glutamine abolished N-linked glycosylation of the V2 receptor (N22Q-V2R), without altering its function or level of expression.** The N22Q-V2R expressed in transfected cell migrated in denaturing ..

- MutationTagger extracts mutations from literature abstracts on a sentence-by-sentence basis
 - textual descriptions (rules)
e.g. *Asparagine substituted with Glutamine at position 22, Mutagenesis of Asparagine 22 to Glutamine*
 - 1- or 3 letter representations (regular expressions)
e.g. *N22Q, N22->Glu, Asn(22)->Glu*
- tested on set of abstracts provided by Caporaso et al. *Bioinformatics* 2007 **23**:1862-1865
 - 88% F-measure

Gene - mutation relations

- identification of genes in text - OK (~ 81%)
- identification of mutations in text - OK (~ 88%)
- challenge: association of mutation to protein
 - > 1 mutation and > 1 protein mentioned in abstract
 - purely based on co-occurrence - high recall, low precision
 - based on distance - low recall, high precision?
- need for a method:
 - identify true relations
 - reduce false positives from individual tasks
- utilising amino acid sequences

Sequence checks

From PMID: 10770218:

.. nephrogenic diabetes insipidus (NDI) is a rare inherited disorder characterized by the excretion of abnormal large volumes of diluted urine mainly caused by mutations in the V2 vasopressin receptor (**AVPR2**) gene. By screening NDI patients for mutations within the **AVPR2** gene we have identified three novel (**I46K**, **F105V**, **I130F**) and four recurrent (**D85N**, **R106C**, **R113W**, **Q225X**) mutations. In addition, a recurrent missense mutation (**A147T**) within the **aquaporin-2** gene was identified in a female patient with autosomal recessive NDI associated with sensorineural deafness. ...

Sequence checks

From PMID: 10770218:

.. nephrogenic diabetes insipidus (NDI) is a rare inherited disorder characterized by the excretion of abnormal large volumes of diluted urine mainly caused by mutations in the V2 vasopressin receptor (**AVPR2**) gene. By screening NDI patients for mutations within the **AVPR2** gene we have identified three novel (**I46K**, **F105V**, **I130F**) and four recurrent (**D85N**, **R106C**, **R113W**, **Q225X**) mutations. In addition, a recurrent missense mutation (**A147T**) within the **aquaporin-2** gene was identified in a female patient with autosomal recessive NDI associated with sensorineural deafness. ...

Sequence checks

From PMID: 10770218:

.. nephrogenic diabetes insipidus (NDI) is a rare inherited disorder characterized by the excretion of abnormal large volumes of diluted urine mainly caused by mutations in the V2 vasopressin receptor (**AVPR2**) gene. By screening NDI patients for mutations within the **AVPR2** gene we have identified three novel (**I46K**, **F105V**, **I130F**) and four recurrent (**D85N**, **R106C**, **R113W**, **Q225X**) mutations. In addition, a recurrent missense mutation (**A147T**) within the **aquaporin-2** gene was identified in a female patient with autosomal recessive NDI associated with sensorineural deafness. ...

Sequence checks

From PMID: 10770218:

.. nephrogenic diabetes insipidus (NDI) is a rare inherited disorder characterized by the excretion of abnormal large volumes of diluted urine mainly caused by mutations in the V2 vasopressin receptor (**AVPR2**) gene. By screening NDI patients for mutations within the **AVPR2** gene we have identified three novel (**I46K**, **F105V**, **I130F**) and four recurrent (**D85N**, **R106C**, **R113W**, **Q225X**) mutations. In addition, a recurrent missense mutation (**A147T**) within the **aquaporin-2** gene was identified in a female patient with autosomal recessive NDI associated with sensorineural deafness. ...

Identified genes:

			121	130	140	147150
359	aquaporin-2	SWISS:P41181	LSNSTTAGQ	AVTVELFLTLQLVLCIF	ASTD	
		SWISS:Q6FGT3	LSNSTTAGQ	AVTVELFLTLQLVLCIF	ASTD	
554	AVPR2	SWISS:O43192	VGMYASSY	ILAMTLDRHRAICR	PMLAYRH	
		SWISS:P30518	VGMYASSY	ILAMTLDRHRAICR	PMLAYRH	
		SWISS:Q9UCV9	WVSTSPAV			




Identified mutations:

I130F: AVPR2(544) -> right
A147T: aquaporin-2(359)-> right, AVPR2(544) -> wrong

Refinement of gene identification

- consider only valid combinations
- re-ranking of candidates

*"An analogous interaction may stabilize the developing positive charge on the Trp-191 radical of the wild-type enzyme. While the oxidation of imidazoles by the ferryl intermediate of **W191G** was neither expected nor observed, this study has defined the structural determinants for small molecule binding to an artificially created cavity near a heme center which is capable of generating oxidized species at a potential of over 1 V, and these results will guide future attempts for novel substrate oxidation by **CCP**"*

CCP [Human]  GeneID: 1421 Seq: 1-174	Ccp1 [Mouse]  GeneID: 67269 Seq: ..TNS V NSV...	CCP1 [Yeast]  GeneID: 853940 Seq: ..EGP W GAA..
---	---	---

Mutation identification and database

Mutation identification

- mutation mining on test set by Caporaso (182 abstracts)
 - F-measure 88%
- mutation - protein relations on subset of 22 abstracts for supported species
 - 17 out of 22 correct mutation - protein combinations (77%)
 - after re-ranking of candidates 20 out of 22 (91%)

Mutation database

- ~ 115,000 mutations
- sequence checked
- for human, mouse, yeast, rat, fruit fly, *H. pylori*, *S. pombe*, *C. elegans*, *A. thaliana*, *D. rerio*

Application: Stability of membrane proteins

- point mutation can influence stability
- simple energy model based on pair potentials
 - calculated on 20 membrane proteins
 - energy for bringing amino acid i from outside to inside
 - $n_{i,in}$: number of inside occurrences
 - $n_{i,out}$: number of outside occurrences

F Dressel et al., *Proc. of the CBSB08 conference*

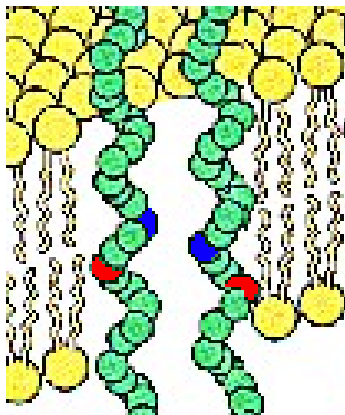
- for 5 G-protein-coupled receptors
rhodopsins: light harvesting
 - mutant phenotypes in literature
 - compare prediction vs descriptions

Application: Stability of membrane proteins

- point mutation can influence stability
- simple energy model based on pair potentials
 - calculated on 20 membrane proteins
 - energy for bringing amino acid i from outside to inside
 - $n_{i,in}$: number of inside occurrences
 - $n_{i,out}$: number of outside occurrences

F Dressel et al., *Proc. of the CBSB08 conference*

- for 5 G-protein-coupled receptors
rhodopsins: light harvesting
 - mutant phenotypes in literature
 - compare prediction vs descriptions

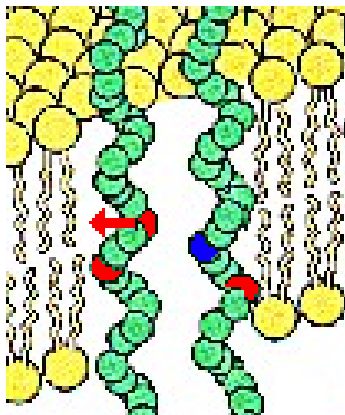


Application: Stability of membrane proteins

- point mutation can influence stability
- simple energy model based on pair potentials
 - calculated on 20 membrane proteins
 - energy for bringing amino acid i from outside to inside
 - $n_{i,in}$: number of inside occurrences
 - $n_{i,out}$: number of outside occurrences

F Dressel et al., *Proc. of the CBSB08 conference*

- for 5 G-protein-coupled receptors
rhodopsins: light harvesting
 - mutant phenotypes in literature
 - compare prediction vs descriptions



Predicting effects of mutations based on sequence

- articles via NCBI utilities
 - PDB
 - SwissProt
 - gene/ protein name + synonyms
- MutationTagger
- sequence checks
- extracted mutations export as HTML
- manual comparison with pair potential based approach
- 25 out of 35 in compliance

Protein name	Mutation in literature	Effect in literature	Stability change	Compliance
Bacteriorhodopsin	G113Q	destabilized	s. destab.	yes
	G113L	destabilized	destab.	yes
	G116Q	destabilized	s. destab.	yes
	G116L	destabilized	destab.	yes
	I117F	destabilized	s. destab.	yes
	I117A	destabilized	stab.	no
	M145F	still active	destab.	no
Halo-rhodopsin	H95A	destabilized	s. destab.	yes
	H95R	destabilized	s. stab.	no
	R108Q	not functional	s. destab.	yes
	T203V	less active	destab.	yes
Rhodopsin	T93P	misfolded	destab.	yes
	T94I	nightblindness	destab.	yes
	C110F	r. pigmentosa	destab.	yes
	C110Y	r. pigmentosa	destab.	yes
	C110A	r. pigmentosa	s. destab.	yes
	E122Q	still active	s. destab.	no
	E122D	still active	s. destab.	no
	E122A		s. destab.	
	E122R	no retinal binding	s. destab.	yes
	C185A	wrong disulfide	s. destab.	yes
	G188R	misfolding	s. destab.	yes
	S186A	incr. activation energy	s. destab.	yes
	C187Y	r. pigmentosa	destab.	yes
	C187A	r. pigmentosa	s. destab.	yes
	N310C	less activity	s. destab.	yes
M317C	less activity	s. destab.	yes	
Antipporter	A130C		s. stab.	
	D133A	not functional	s. destab.	yes
	H225P	less activity	destab.	yes
	H225C	less activity	none	no
	G303C	not functional	s. destab.	yes
Aquaporin	N42A	still active	s. destab.	no
	A73M	not functional	s. stab.	no
	Y186F	conduct water	s. destab.	
	Y186A	no water conductance	stab.	
	Y186N	no water conductance	stab.	
	C189M	less activity	s. stab.	no
	C189S	still active	s. stab.	yes
	H209A	still active	s. destab.	no

Example

Protein name	Mutation in literature	Effect in literature	Stability change	Compliance
Rhodospin	T93P	misfolded	destab.	yes

- mutation T93P in bovine rhodopsin
- reported to lead to conformational change
Kono et al. Biochemistry 2005 44(2):799-804
- compare solvation energies
 - threonine: negative
 - proline: positive
 - destabilizing effect
- in compliance with literature
- prediction is correct

Conclusion

- retrieve protein point mutations
 - flexible for any given gene
 - species specific from whole PubMed
- analysis pipeline for biomedical applications
 - stability of membrane proteins
- population of database for mutations from literature
 - proteins - interactions - diseases
- protein specificity through sequence checks
- improvement of gene tagging

Acknowledgments

Bioinformatics group @ Biotec - TU Dresden

- Group Leader
 - Michael Schroeder
- Text Mining
 - Conrad Plake
 - Jörg Hakenberg (Arizona)
 - .. amongst others
- Transmembrane Proteins
 - Frank Dressel
 - Annalisa Marsico
 - Anne Tuukkanen
 - Dirk Labudde



Mutation tagging with gene identifiers applied to membrane protein stability prediction

Rainer Winnenburg, Conrad Plake, and Michael Schroeder

AIMM workshop @ ECCB'08

September 22, 2008

Thank you for your attention!