title: Predict effect of mutation on function
short title: pp2_snp1
lecture: Protein Prediction 2 - Protein function
TUM winter 2013/2014
Announcements

 Videos: SciVe / www.rostlab.org

 THANKS:

 Tim Karl + Jona Reeb

 Special lectures:

 • Oct 29: Tobias Hamp
 • Nov 21: Tanya Goldberg
 • Nov 28: Arthur Dong
 • Dec 03+05: Marco De Vivo/Marco Punta
 • Dec 17+19: Andrea Schafferhans

 No lecture:

 • Oct 10 Thu (Reformation)
 • Nov 12 Tue (Student assembly)
 • Dec 12 Thu (TUM Dies Academicus)

 LAST lecture: Jan 30

 Examen: Feb 4 (CompSci) and Feb 6 (CompBio) - likely this room

 • Makeup: Apr 9 - morning

 CONTACT: Marlena Drabik assistant@rostlab.org
V.
Predict effect of mutations
V.1 SNP effect: Intro
Life is diverse
Evolution

Charles Darwin: Origin of Species, Nov 24, 1859
Evolution: some things change

Charles Darwin: Origin of Species, Nov 24, 1859
Evolution: others stay the same

© Wikipedia

Alfred Russell Wallace & Charles Darwin: 1858

On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection

Two papers read (by Secretary John Joseph Bennett) to the Linnean Society (London) Jul 1, 1858
“Darwin-Wallace Theory”

© Burkhard Rost (TUM Munich)
Predict protein function
Big changes may not matter!

Don't know region

Sequence identity implies structural similarity!

Percentage sequence identity

Number of residues aligned

Distance from curve = +10

Distance from curve = -10

C Sander & R Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94

Tuesday January 21, 2014
SNPs are changes of one single nucleotide / letter

SNPs are changes of one single nucleotide / letter

Wikipedia
Map genotype to phenotype

© US Dept Health & Human Services
www.hhs.gov/ohrp/sachrp/mtgings/mtg07-08/present/koenig.html
Discover diversity and movement

“Bushmen ... more different from each other than ... European and Asian”

“migration from Siberia into New World ... 5,500 years ago, independent of ... Native Americans and Inuit”
Who Were the Denisovans?

Science 26 August 2011:
vol. 333 no. 6046 1084–1087

Genetic history of an archaic hominin group from Denisova Cave in Siberia

David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin & Svante Pääbo

Affiliations | Contributions | Corresponding authors

Nature 468, 1053–1060 (23 December 2010) | doi:10.1038/nature09710
Received 15 August 2010 | Accepted 30 November 2010 | Published online 22 December 2010

Abstract

Abstract • Introduction • DNA sequence determination • Human DNA contamination estimates • Ancestral features and duplications • Relationship to Neanderthals and modern humans • A Neanderthal-specific bottleneck • No Denisovan gene flow into all Eurasians • Denisovan gene flow into the ancestors of Melanesians • A model of population history • Discordance of mtDNA and nuclear histories • A tooth from Denisova Cave • Morphology of the Denisova molar • Stratigraphy and dating • Discussion • Methods • Accession codes • References • Acknowledgements • Author information • Supplementary information • Comments
non-synonymous SNP = nsSNP

Mutation

SNP – single nucleotide polymorphism
Coding region SNP
nsSNP – non-synonymous SNP

DNA
ACGGACCGGACTCTCATATATACCTACTCCTTGATGGGCTCCGAAATATG

Non-coding region
Coding region

Protein: MGSEIC
non-synonymous SNP = nsSNP

Mutation

SNP – single nucleotide polymorphism
Coding region SNP
nsSNP – non-synonymous SNP

DNA
ACGGACGGACCTCTCATATACTACTCCTTGATGGGCTCCGAAATATG

Non-coding region

Coding region

Protein: MGSEIO
Lingo

- change of amino acid: non-synonymous (ns)
- occurrence > T% in population: SNP

here:

ANY amino acid change
-> nsSNP (or SNP)
SNP risk classification

Non-sense
Splicing regulation
Mis-sense/non-synonymous (nsSNPs)
Regulatory region
Post-transcriptional regulation
Untranslated / up- or downstream

nsSNPs importantly involved in disease

Figs. 2 - 1a
V.2 nsSNP effect:
Other methods
SIFT: Sorting Intolerant From Tolerant
Fred Hutchinson Cancer Center, Seattle
HHMI (Howard Hughes Medical Institute)

papers:
- >300 papers (Nov 2011)
- 3 >1,000 citations (end 2011)
- 72 over 100
- H-index 83 (ISI Nov 2011)

Paradigm changes
- gene in gene - in intron (1986)
- histones NOT only in octamers (2004)
- DNA-methylation in histones: H2.AZ in histone spool promotes gene expression (2008): NOT DNA-methylation shuts off genes (important for cancer drug development)
Pauline Crystal Ng

☐ Genome Institute of Singapore (Senior Investigator)

☐ CV:
  • BS: Caltech
  • PhD (Bioengineering): Univ of Washington, Fred Hutchinson, Seattle (with Steven Henikoff)
  • Postdoc: Illumina
  • Assistant Professor: J Craig Venter Institute, Seattle

☐ papers:
  • >20 papers (Nov 2011)
  • 3 >300 citations (end 2011)
  • common name: >300 papers, most of the top cited ones from HER!
Test Set: 3 proteins, ~6500 mutants

PolyPhen: Polymorphism Phenotyping
Shamil R Sunyaev

☐ Harvard Medical School
   (Assistant Professor)

☐ CV:
   • PhD: Moscow Inst. of Physics & Technology EMBL
   • Postdoc: EMBL

☐ papers:
   • >50 papers (Nov 2011)
   • 3 >100 citations (end 2011)
PolyPhen

Test Set: 1551 SWISS-PROT & 440 cross-species variants

Train/Test Set: 3768 HGMD variants and 2309 cross-species variants

nsSNP effects: some *in silico* methods

- **SIFT**  
  PC Ng & S Henikoff (2003) NAR 31:3812-14

  Sequence:
  VHLTP EKSA VTALWGKVNV
  DEVGGEALGR LLVVYPWTQR
  FFESFGDLST PDAVMGNPKV
  KAHGKKVLGA
  Mutant: E6V

- **PolyPhen**  

  Sequence:
  VHLTP EKSA VTALWGKVNV
  DEVGGEALGR LLVVYPWTQR
  FFESFGDLST PDAVMGNPKV
  KAHGKKVLGA
  Mutant: E6V

- **SNPs3D**  
  P Yue, Z Li & J Moult (2005) JMB 353:459-63

  Sequence:
  VHLTP EKSA VTALWGKVNV
  DEVGGEALGR LLVVYPWTQR

© Burkhard Rost (TUM Munich)
SDM (Site Directed Mutator)

Thermodynamic cycle

CM Topham, N Srinivasan & TL Blundell (1997) Protein Eng. 10, 7-21

© http://mordred.bioc.cam.ac.uk/~sdm/sdm_theory.php
CM Topham, N Srinivasan & TL Blundell (1997) Protein Eng. 10, 7-21

© http://mordred.bioc.cam.ac.uk/~sdm/sdm_theory.php
SDM (Site Directed Mutator)

Thermodynamic cycle

CM Topham, N Srinivasan & TL Blundell (1997) Protein Eng. 10, 7-21

© http://mordred.bioc.cam.ac.uk/~sdm/sdm_theory.php
PoPMuSiC – Prediction of Protein Mutant Stability Changes

Changes in binding energy learned by neural networks from statistical potentials

Y Dehouck et al & M Rooman (2009) Bioinformatics 25, 2537-43

© http://babylone.ulb.ac.be/popmusic

Marianne Rooman

© Burkhard Rost (TUM Munich)
Changes in binding energy learned by SVM from ProTherm

Emidio Capriotti, Piero Fariselli & Rita Casadio (2005) NAR 33, W306-10

© http://folding.biofold.org/i-mutant/i-mutant2.0.html
AutoMute - AUTOmated server for predicting ... functional consequences of amino acid MUTations in protEins

Changes in binding energy learned by Machine Learning considering 3D environment


© http://proteins.gmu.edu/automute/
## Prediction Performance Comparison

**SDM, Crescendo, (PICCOLO, CREDO and BIPA), SIFT, MUpro, MAPP and I-Mutant2.0**

<table>
<thead>
<tr>
<th>Method</th>
<th>TP</th>
<th>FP</th>
<th>TN</th>
<th>FN</th>
<th>Sum</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDM</td>
<td>535</td>
<td>274</td>
<td>4594</td>
<td>3025</td>
<td>8428</td>
<td>15.03</td>
<td>94.37</td>
<td>60.86</td>
</tr>
<tr>
<td>Crescendo</td>
<td>186</td>
<td>126</td>
<td>4482</td>
<td>3695</td>
<td>8489</td>
<td>4.79</td>
<td>97.27</td>
<td>54.99</td>
</tr>
<tr>
<td>PICCOLO</td>
<td>220</td>
<td>257</td>
<td>4920</td>
<td>3746</td>
<td>9143</td>
<td>5.55</td>
<td>95.04</td>
<td>56.22</td>
</tr>
<tr>
<td>CREDO</td>
<td>380</td>
<td>413</td>
<td>4764</td>
<td>3586</td>
<td>9143</td>
<td>9.58</td>
<td>92.02</td>
<td>56.26</td>
</tr>
<tr>
<td>BIPA</td>
<td>80</td>
<td>20</td>
<td>5157</td>
<td>3886</td>
<td>9143</td>
<td>2.02</td>
<td>99.61</td>
<td>57.28</td>
</tr>
<tr>
<td>COMBINED</td>
<td>1252</td>
<td>984</td>
<td>4193</td>
<td>2714</td>
<td>9143</td>
<td>31.57</td>
<td>80.99</td>
<td>59.55</td>
</tr>
<tr>
<td>SIFT</td>
<td>2709</td>
<td>2071</td>
<td>3011</td>
<td>1092</td>
<td>8883</td>
<td>71.27</td>
<td>59.25</td>
<td>64.39</td>
</tr>
<tr>
<td>MAPP</td>
<td>2659</td>
<td>1642</td>
<td>2395</td>
<td>1065</td>
<td>7761</td>
<td>71.40</td>
<td>59.33</td>
<td>65.12</td>
</tr>
<tr>
<td>I-Mutant2.0</td>
<td>1485</td>
<td>1677</td>
<td>2189</td>
<td>1061</td>
<td>6412</td>
<td>58.33</td>
<td>56.62</td>
<td>57.30</td>
</tr>
<tr>
<td>MUpro</td>
<td>175</td>
<td>146</td>
<td>5031</td>
<td>3791</td>
<td>9143</td>
<td>4.41</td>
<td>97.18</td>
<td>56.94</td>
</tr>
</tbody>
</table>

Table 1. TP= True Positives, FP= False Positives, TN= True Negatives. TP/FP/TN/FN are numbers of unique mutations. The Sum column shows the number of times the method succeeded and an observation was possible and therefore reflects the robustness of the method. Sensitivity, Specificity and Accuracy defined in text.
V.3 nsSNP effect:

Data
Misfunction/neutral
ENTRY A000006 - Variant 2616650
CROSS-REFERENCE MC4R_HUMAN
PROTEIN Melanocortin 4 receptor (MC4R);
CHANGE-POINT Asn 62 Ser (homozygous)
DISEASE In obesity
FUNCTION Responsiveness to alpha-MSH [-]

ENTRY A000006 - Variant 2616650
AUTHORS Farooqi I.S., Yeo G.S.H., Keogh J.M., Aminian S., Jebb S.A., Butler G., Cheetham T. & O'Rahilly S.
CROSS-REFERENCE MC4R_HUMAN
PROTEIN Melanocortin 4 receptor (MC4R);
CHANGE-POINT Asn 62 Ser (homozygous)
DISEASE In obesity
FUNCTION Responsiveness to alpha-MSH [–]

Effect: 40,641
Neutral: 14,334

ENTRY A000006 - Variant 2616650
AUTHORS Farooqi I.S., Yeo G.S.H., Keogh J.M., Aminian S., Jebb S.A., Butler G., Cheetham T. & O'Rahilly S.
CROSS-REFERENCE MC4R_HUMAN
PROTEIN Melanocortin 4 receptor (MC4R);
CHANGE-POINT Asn 62 Ser (homozygous)
DISEASE In obesity
FUNCTION Responsiveness to alpha-MSH [-]

Effect: 40,641
Neutral: 14,334

Machine-Learning handles imbalance?

- Effect: 26%
- Neutral: 74%
SNAP data: neutral

EC# = general_class . acts_on_class . further_class_spec . spec_by_substrate_class
3.1.3.48 \rightarrow hydrolase . on ester bonds . phosphoric monoester cmpnds . PTP-phosphotase

Same EC# = Same Function

Non-neutral: 40,641
Neutral: 14,334

Non-neutral: 26,840
Neutral: 26,840

Total
Non-neutral: 40,641
Neutral: 41,174

81,815 mutants
6,821 proteins

Y Bromberg & B Rost 2007 NAR 35:3823-35
### SNAP data: by accessibility

<table>
<thead>
<tr>
<th>Accessory</th>
<th>Proteins</th>
<th>Effect</th>
<th>Neutral</th>
<th>Total</th>
<th>Ratio neutral to non-neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>6413</td>
<td>39987</td>
<td>40830</td>
<td>80817</td>
<td>1.02</td>
</tr>
<tr>
<td><strong>Buried</strong></td>
<td>5144</td>
<td>19741</td>
<td>14800</td>
<td>34541</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Intermediate</strong></td>
<td>4841</td>
<td>12285</td>
<td>13073</td>
<td>25358</td>
<td>1.06</td>
</tr>
<tr>
<td><strong>Exposed</strong></td>
<td>4150</td>
<td>7961</td>
<td>12957</td>
<td>20918</td>
<td>1.62</td>
</tr>
</tbody>
</table>
V.4 nsSNP effect: SNAP predictions
Yana Bromberg, Rutgers University
SNAP: input features

**SNAP**
- Biochemical characteristics
- Alignment profiles
- Probability of residue triplets
- Pfam domains
- Solvent accessibility
- Secondary structure
- Residue flexibility

**SNAP^annotated**
- SWISS-PROT annotations
- SIFT predictions
SNAP: neural network
SNAP: neural network
SNAP: neural network

<table>
<thead>
<tr>
<th>SNP effect</th>
<th>Node score</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral</td>
<td>min</td>
</tr>
<tr>
<td>non-neut</td>
<td>max</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNP effect</th>
<th>Node score</th>
</tr>
</thead>
<tbody>
<tr>
<td>neutral</td>
<td>max</td>
</tr>
<tr>
<td>non-neut</td>
<td>min</td>
</tr>
</tbody>
</table>
SNAP: neural network

SNP effect | Node score
-neut min
neutral max

Score: -100 ≤ S ≤ 100
## Performance comparison

### Overall Two-State Accuracy (Q2)

<table>
<thead>
<tr>
<th>Method</th>
<th>LacI Repressor</th>
<th>Lysozyme</th>
<th>HIV-1 protease</th>
<th>Melanocortin-4 receptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>69.4</td>
<td>67.6</td>
<td>78.3</td>
<td>57.8</td>
</tr>
<tr>
<td>PolyPhen</td>
<td>68.7</td>
<td>57.9</td>
<td>***</td>
<td>51.1</td>
</tr>
<tr>
<td>SNAP</td>
<td>70.7</td>
<td>70.0</td>
<td>68.5</td>
<td>71.1</td>
</tr>
<tr>
<td>SNAP\text{annotated}</td>
<td>72.7</td>
<td>73.2</td>
<td>72.3</td>
<td>80.0</td>
</tr>
<tr>
<td>SNPs3D</td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>62.2</td>
</tr>
</tbody>
</table>

\[ Q2 = 100 \times \frac{\text{# correct predictions}}{\text{total # of predictions}} \]
SNAP performs well

80,000 mutants w/known effects on function

Random
SIFT
PolyPhen
SNAP

© Yana Bromberg, 2010 Columbia University

Y Bromberg & B Rost 2007 NAR 35:3823-35
SNAP clearly best for subtle cases

<table>
<thead>
<tr>
<th>PMD/EC data</th>
<th>Unknown</th>
<th>Accuracy effect</th>
<th>Coverage effect</th>
<th>Accuracy neutral</th>
<th>Coverage neutral</th>
<th>Overall two-state accuracy (Q2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>2374 (3%)</td>
<td>79.8±0.6</td>
<td>63.4±1.2</td>
<td>70.1±2.7</td>
<td>84.3±1.2</td>
<td>74.0±1.4</td>
</tr>
<tr>
<td>PolyPhen</td>
<td>1647 (2%)</td>
<td>79.1±0.7</td>
<td>66.9±1.4</td>
<td>71.8±2.7</td>
<td>82.7±1.1</td>
<td>74.9±1.3</td>
</tr>
<tr>
<td>SNAP</td>
<td>0</td>
<td>76.3±0.8</td>
<td>83.3±1.0</td>
<td>82.0±2.4</td>
<td>74.7±2.2</td>
<td>78.9±1.3</td>
</tr>
</tbody>
</table>

Accuracy = 100 * \(\frac{\text{# correct predictions}}{\text{total # of predictions}}\)

Coverage = 100 * \(\frac{\text{# correct predictions}}{\text{total # of observations}}\)
SNAP performs well

80,000 mutants w/known effects on function

Random
SIFT
PolyPhen
SNAP

© Yana Bromberg, 2010 Columbia University

© Burkhard Rost (TUM Munich)
SNAP much better for tough cases

80,000 mutants w/ known effects on function

© Yana Bromberg, 2010 Columbia University

© Burkhard Rost (TUM Munich)
SNAP clearly best for subtle cases

<table>
<thead>
<tr>
<th>PMD/EC data</th>
<th>Unknown</th>
<th>Accuracy effect</th>
<th>Coverage effect</th>
<th>Accuracy neutral</th>
<th>Coverage neutral</th>
<th>Overall two-state accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>2374 (3%)</td>
<td>79.8±0.6</td>
<td>63.4±1.2</td>
<td>70.1±2.7</td>
<td>84.3±1.2</td>
<td>74.0±1.4</td>
</tr>
<tr>
<td>PolyPhen</td>
<td>1647 (2%)</td>
<td>79.1±0.7</td>
<td>66.9±1.4</td>
<td>71.8±2.7</td>
<td>82.7±1.1</td>
<td>74.9±1.3</td>
</tr>
<tr>
<td>SNAP</td>
<td>0</td>
<td>76.3±0.8</td>
<td>83.3±1.0</td>
<td>82.0±2.4</td>
<td>74.7±2.2</td>
<td>78.9±1.3</td>
</tr>
</tbody>
</table>

**“Hard” PMD/EC data**

<table>
<thead>
<tr>
<th>PMD/EC data</th>
<th>Correct classification</th>
<th>Incorrect classification</th>
<th>Overall two-state accuracy (Q2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>7,566</td>
<td>8,675</td>
<td>46.6%</td>
</tr>
<tr>
<td>PolyPhen</td>
<td>7,966</td>
<td>8,275</td>
<td>49.1%</td>
</tr>
<tr>
<td>SNAP</td>
<td>10,124</td>
<td>6,117</td>
<td>62.3%</td>
</tr>
</tbody>
</table>

Accuracy = \(100 \times \frac{\text{# correct predictions}}{\text{total # of predictions}}\)

Coverage = \(100 \times \frac{\text{# correct predictions}}{\text{total # of observations}}\)
SNAP in Pictures

80,000 mutants w/known effects on function

Predict
Features

Accuracy = 100 * \frac{\text{# correct predictions}}{\text{total # of predictions}}

Coverage = 100 * \frac{\text{# correct predictions}}{\text{total # of observations}}

© Yana Bromberg, 2010 Columbia University

Y Bromberg & B Rost 2007 NAR 35:3823-35

© Burkhard Rost (TUM Munich)

Tuesday January 21, 2014
SNAP performance by exposure
SNAP reliability index

Predictions

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.49</td>
<td></td>
</tr>
</tbody>
</table>
SNAP reliability index

![Graph showing cumulative accuracy vs. cumulative percentage predicted with RI≥n for non-neutral and neutral categories.](image-url)
SNAP RI ~ severity of change

Normalized percentage of predictions

Difference between two SNAP output units
## SNAP: examples

<table>
<thead>
<tr>
<th>Gene</th>
<th>nsSNP</th>
<th>Disease</th>
<th>Function</th>
<th>Prediction (RI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HXK4</td>
<td>S131P</td>
<td>Diabetes Mellitus</td>
<td>Significant increase in affinity for ATP</td>
<td>Non-neutral (2)</td>
</tr>
<tr>
<td>PAX6</td>
<td>G64V</td>
<td>Cataract</td>
<td>Reduction of DNA binding activity</td>
<td>Non-neutral (5)</td>
</tr>
<tr>
<td>MC4R</td>
<td>I301T</td>
<td>Obesity</td>
<td>Severe change of basal activity &amp; EC50</td>
<td>Non-neutral (3)</td>
</tr>
<tr>
<td>HXK4</td>
<td>M107T</td>
<td>Polymorphism</td>
<td>Not conclusive</td>
<td>Neutral (0)</td>
</tr>
<tr>
<td>CFTR</td>
<td>P1013L</td>
<td>Cystic Fibrosis</td>
<td>Not conclusive</td>
<td>Non-neutral (5)</td>
</tr>
<tr>
<td>NKX25</td>
<td>A127G</td>
<td>Secundum atrial septal defect</td>
<td>Not conclusive</td>
<td>Neutral (6)</td>
</tr>
<tr>
<td>HBB</td>
<td>R104T</td>
<td>Polymorphism</td>
<td>Not conclusive</td>
<td>Non-neutral (3)</td>
</tr>
<tr>
<td>P53</td>
<td>R337H</td>
<td>Adrenocortical carcinoma</td>
<td>Does not affect transactivation capacity</td>
<td>Non-neutral (3)</td>
</tr>
</tbody>
</table>
Annotations help

![Bar chart showing percentage of SNPs predicted correctly for Annotation, Alignment, and All categories.](Image)
Annotations help, but not often
Crucial sites identified in insulin

http://www.rostlab.org/servers/SNAP/

Y Bromberg & B Rost
2007 NAR 35:3823-35

Y Bromberg G
Yachdav & B Rost
2008 Bioinformatics
15:2397-8

model: SWISS-MODEL
SNAP: glucokinase

V.5 nsSNP effect:
SNAP beyond singles
New directions

- in silico alanine scan
- comprehensive in silico mutagenesis
- prediction of binding hot spots

Y Bromberg & B Rost (2008) Bioinformatics 24: i207-212
Experimental mutagenesis

Targeted Mutagenesis

Ala, Cys, Gly,…

Residue Scan

Ala
In silico mutagenesis

Targeted Mutagenesis

?  

Residue Scan

comprehensive all against 19 non-native
### Prediction and mutagenesis

**MC4R_HUMAN**

MVNSTHRGMHTSLHLWNRSSYRLHSNASESLGKGYSDG
GCYEQLFVSPEVFVTGVISLLLENILVIVAIAKKNKLHSPMY
FFICSLAVADMLVSNSGSEIVITLLNSTDYDAQSFTVNID
NVIDSVICSSLLASICLLSIAVDRYFTIFYALQYHNIMTVKR
VGIIIISCIWAACTVSGILFIYSDSAVIICLITMFFTYMLALMAS
LYVHMFLMARLHIKRIAVLPGTGAIRQGANMKGAIITLTLIG
VFVVCWAPFFHHLIFYISCPCQNPYCFCMSHFNLYLILIMCN
SIIIDPLIYALRSQELRKTKEIICCYPGLCGDSSRY

<table>
<thead>
<tr>
<th>nsSNP</th>
<th>Prediction</th>
<th>Reliability</th>
<th>Exp Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>R7H</td>
<td>Neutral</td>
<td>5</td>
<td>89%</td>
</tr>
<tr>
<td>S30F</td>
<td>Non</td>
<td>4</td>
<td>82%</td>
</tr>
<tr>
<td>E100A</td>
<td>Non</td>
<td>3</td>
<td>78%</td>
</tr>
</tbody>
</table>

R7H, S30F, E100A

---

**Prediction and mutagenesis**

Bromberg & B Rost 2008 *Bioinformatics* 24: i207-212

slide: © Yana Bromberg
Prediction and mutagenesis

>MC4R_HUMAN
MVNSTHRGMHTSLHLWRSSYRLHSNASESLGKGYSDG
GCYEQLFVSPEVFVTLGVISLLENILVIVAIAKNKLHSPMY
FFICSLAVADMLVSVSNGSEETIVITLNNSTDTDAQSFTVNID
NVIDSVICSSLLASICLLLSIAVDRYFTIFYALQYHNIIMTVKR
VGIIIISCIWAACTVSGILFIYSDDSSAVIICLITMFTMLALMAS
LYVHMFLMARLHIKRIAVALPGTGAIORQGANMKGAITALTLIG
VFVVCWAPFFLHLIFYISCPQNPYCVCMSHFNLYLILIMCN
SIIDPLIYALRSQELRKTKEIICCYPPLGLCDLSSRY

<table>
<thead>
<tr>
<th>nsSNP</th>
<th>Prediction</th>
<th>Reliability</th>
<th>Exp Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>R7H</td>
<td>Neutral</td>
<td>5</td>
<td>89%</td>
</tr>
<tr>
<td>S30F</td>
<td>Non</td>
<td>4</td>
<td>82%</td>
</tr>
<tr>
<td>E100A</td>
<td>Non</td>
<td>3</td>
<td>78%</td>
</tr>
</tbody>
</table>

R7H, S30F, E100A

ALL MUTANTS
(19 substitutions per position)

Y Bromberg & B Rost 2008 *Bioinformatics* 24: i207-212
### Prediction and mutagenesis

**MC4R_HUMAN**

MVNSTHRGMLTHSLHLWNRSYRLHSSASESLGKGYSDG
GCYEQLFVSPEVFVTGLGVISLLENILVIVAIAKNKNLHSPMY
FFICSLAVADMLVSNSNGSETIVITLNNSTDTDASSFTVNIID
NVIDSVICSSLLASICSOLLSSIAVDRYFTIFYALQYHNIMTVKR
VGIIISCIWAACTVSGILFIYSDSASVIICLITMFFTLMALMAS
LYVHMFLMARLHIKRIAVLPGTAIRQGANGMKGAILTLILIG
VFVVCWAPFFLHLFYISCPQNYPVCFSNFNYLILICMC
SIIDPLIYALRSQELRKTFKEICcyplgglcdlssry

<table>
<thead>
<tr>
<th>nsSNP</th>
<th>Prediction</th>
<th>Reliability</th>
<th>Exp Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>R7H</td>
<td>Neutral</td>
<td>5</td>
<td>89%</td>
</tr>
<tr>
<td>S30F</td>
<td>Non</td>
<td>4</td>
<td>82%</td>
</tr>
<tr>
<td>E100A</td>
<td>Non</td>
<td>3</td>
<td>78%</td>
</tr>
</tbody>
</table>

**ALL MUTANTS (19 substitutions per position)**

**ALL PREDICTIONS (19 substitutions per position)**

Y Bromberg & B Rost *2008* *Bioinformatics* 24: i207-212

slide: © Yana Bromberg
In silico “alanine” scan

Correlation of average over 19 possible SNAP predictions per location to single residue scores

Y Bromberg & B Rost 2008 Bioinformatics 24: i207-212
Mutation effects reveal functional units
Protein-protein hot spot binding

![Graph showing distribution of ddG range for non-neutral and neutral samples.](image)

- Non-neutral
- Neutral

Number of samples in given ddG range

ddG range (kCal/mol)
Protein-protein hot spot binding
Predict binding hotspots

Y Bromberg & B Rost (2008) Bioinformatics 24: i207-212
Important residues in binding sites

Y Bromberg & B Rost 2007 NAR 35:3823-35
Y Bromberg G Yachdav & B Rost 2008 Bioinformatics 15:2397-8
Functional residues

Mean over non-native  Conservation  Other scoring
In silico mutagenesis

Experimental

LacI repressor from E. coli
4011 mutants (12-13 substitutions/residue)
SNAP prediction accuracy: ~73%

Experimental: P Markiewicz et al. 1994 JMB 240:421-33
Y Bromberg & B Rost unpublished

© Yana Bromberg, 2010 Columbia University
© Burkhard Rost (TUM Munich)
**In silico mutagenesis**

**Experimental**

DNA Binding Domain

- $I^{-}$$^{1^{rd}}$
- $I^{+}$
- $I^{-}$

**Predicted (SNAP)**

LacI repressor from E. coli
4011 mutants (12-13 substitutions/residue)
SNAP prediction accuracy: ~73%

Experimental: P Markiewicz et al. 1994 JMB 240:421-33
Y Bromberg & B Rost unpublished

© Yana Bromberg, 2010 Columbia University

© Burkhard Rost (TUM Munich)
Melanocortin receptor (MC4R)

Y Bromberg et al. 2009 FASEB 9:3059-69
Differential view on 2 similar receptors

Y Bromberg et al. 2009 FASEB 9:3059-69
V.6 nsSNP effect: Structure
Predict effect of change upon structure
regular secondary structure withstands mutation, disorder does not
Evolution under constraints

Sequence identity implies structural similarity!

Don't know region

Distance from curve = +10

Distance from curve = -10

C Sander & R Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94
Christian Schaefer
Regular secondary structure sustains random mutations
Astonishing mutations: helix<->strand

C Schaefer & B Rost 2010 *Bioinformatics* 26:625-31
Disordered regions


© Burkhard Rost (TUM Munich)
Long disorder disappears

C. Schaefer & B Rost 2010 Bioinformatics 26:625-31

© Burkhard Rost (TUM Munich)
Long disorder disappears

C Schaefer & B Rost 2010 Bioinformatics 26:625-31

© Burkhard Rost (TUM Munich)
Fatty Acid Synthase
Predict impact of nsSNP upon protein structure

Predict change from sequence alone

- Input: sequence: MSVKELEDKVEELLSKNYHLENEVARLKKLVGER
  mutation: E7M
- Outcome: change / no-change
3D fragments to learn from

[Diagram with text: chain A, chain B, fragment A, fragment B, LKKEN, LKCEN, strctl. similarity measure: RMSD]
Root mean square displacement (rmsd)

\[ \text{rmsd}(A,B) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} d_i^2} \]

- **Step 1:** find corresponding points in proteins A and B
- **d(i)** are the distances between all corresponding points (typic: \(C_{\alpha}\), all atoms)

Two pairs of 29mers

RMSD(Cα)
0.013 nm
(1.3 Å)

RMSD(Cα)
0.146 nm
(14.6 Å)
Separation of RMSD distributions

- **Background**
- **Machine learning**
- **Real world application**

**RMSD $C_\alpha$ [Å]**

- Randoms
- One central mismatch
- Identicals

**5mers**
Let’s prove statistics meaningless and develop a prediction method
Prediction method in different light

Application to mutagenesis sets

Two different points of view

- Stability: ProTherm dataset
- Function: SNAP dataset

Outcome: Predicted probability of having structural effect
V.7 nsSNP effect:

From here
Molecular Dynamics (MD) can be very powerful
Large-scale protein flexibility analysis of single nucleotide polymorphisms using molecular dynamics simulations
Molecular dynamics of SNP in Gaucher disease

Marc Offman

Fig. 1:
Marc N Offman, M Krol, B Rost, I Silman, JL Sussman & AH Futerman (2011)
Validation of a molecular dynamics protein structure PREDICTION:
Comparison of an MD model with the X-ray structure of the N370S acid-β-glucosidase mutant that causes Gaucher disease.
PEDS in press.
MD2: new SNPs causing Parkinson’s Disease

Fig. 2:
A Zimprich et al. & T Meitinger & TM Strom (2011)
Exome sequencing reveals mutations in the retromer protein VPS35 as cause for Parkinson’s disease.
Am J Hum Genet, 89: 168-75

© Burkhard Rost (TUM Munich)
Sets of SNPs:
human - Neandertal - chimp
Homo sapiens vs. Homo neanderthalensis

Artist's rendering of Neandertal man, from Neandertal museum in Mettmann, All rights reserved. Germany Copyright: Johannes Krause, Max Planck Institute for Evolutionary Anthropology, Leipzig.

<table>
<thead>
<tr>
<th>SNAP results</th>
<th>H. sapiens</th>
<th>H. neanderthalensis</th>
<th>P. troglodytes</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>41</td>
</tr>
<tr>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>5</td>
</tr>
<tr>
<td>no effect</td>
<td>effect</td>
<td>effect</td>
<td>no effect</td>
<td>13</td>
</tr>
<tr>
<td>no effect</td>
<td>effect</td>
<td>effect</td>
<td>effect</td>
<td>2</td>
</tr>
<tr>
<td>effect</td>
<td>no effect</td>
<td>no effect</td>
<td>no effect</td>
<td>5</td>
</tr>
<tr>
<td>effect</td>
<td>effect</td>
<td>no effect</td>
<td>no effect</td>
<td>1</td>
</tr>
<tr>
<td>effect</td>
<td>no effect</td>
<td>effect</td>
<td>effect</td>
<td>2</td>
</tr>
<tr>
<td>effect</td>
<td>effect</td>
<td>effect</td>
<td>effect</td>
<td>1</td>
</tr>
</tbody>
</table>

© Burkhard Rost (TUM Munich)
Data set (MPI Leipzig, S Paabo & J Kelso):
78 nsSNPs such that
• fixed in modern human (1000 genomes)
• ancestral in Neandertal & chimp

put differently:
population of modern human
all do NOT have nsSNP but
Neandertal & chimp DO!
Map to human PPIs  (protein protein interactions)

- 78 SNPs in 69 proteins
- only 27 (of 69) map to network (45K PPIs/10K prot)
- Surprisingly few; few pathways
- Stark contrast to other sets (e.g. human viral targets)
  -> less well-studied proteins, may be downstream effectors
Although they are isolated, they still cluster together

SNP proteins are close to one another on the human protein network

Coherent functions and common regulations

Arthur Dong
Particular challenges for human
Protein-protein interactions = Physical interactions NOT associations

HIV gp120 / CD4 / FAB

Networks: pathogen interactions differ

virus adopts cellular interaction features upon invasion

© Arthur Dong
now TUM

Uetz, Dong et al 2006 Science 311: 239-42
Disordered regions

Avner Schlessinger
now: UCSF
Different methods find different proteins

**NORS:** J Liu, H Tan & B Rost 2002 *J Mol Biol* 322:53-64  
**PROFbval:** A Schlessinger & B Rost 2005 *Proteins* 61: 115-126  
**UCon:** A Schlessinger, M Punta & B Rost 2007 *Bioinformatics* 21:2376-84  
**MD:** A Schlessinger & B Rost 2009 *PLoS One*, 4: doi10.1371

**Ucon**
*(contacts)*

**NORSnet (loopy)**

Max transcription factor (TANZ)

Capsid protein from cricket paralysis virus (1b35_C)
Different methods find different proteins


*Ucon* (contacts)

*Max transcription factor (TANZ)*

*Capsid protein from cricket paralysis virus (1b35_C)*
Eukaryotes dominate disorder (4-10x)

A Schlessinger et al & B Rost 2011 *Curr Opin Struc Biol* 21:412-8
Connect, extend, make available
SNP pipeline

Andrea Schafferhans

Laszlo Kajan

Shaila Roessle

Marc Offman

© Burkhard Rost (TUM Munich) 120 /122
Pre-computed annotations 4 human

Maximilian Hecht
Alice Meier
Peter Hoenigschmid

Yana Bromberg

Chris Schaefer
Disease annotations

C Schaefer et al. (2011) Bioinformatics, submitted
Lecture plan (PP2 function)

01: 2013/10/15: no lecture
02: 2013/10/17: welcome: who we are
03: 2013/10/22: Intro - function 1: concepts
04: 2013/10/24: Intro - function 2: homology
05: 2013/10/29: Tobias Hamp: Homology-based prediction of function
06: 2013/10/31: no lecture Reformation
07: 2013/11/05: Tobias Hamp: Homology-based prediction of function 2
08: 2013/11/07: Intro - function 3: motifs
09: 2013/11/12: no lecture: SVV (student reps)
10: 2013/11/14: Localization 1
11: 2013/11/19: Localization 2
12: 2013/11/21: Localization 3 - Tatyana Goldberg
14: 2013/11/28: Localization 4 - Tatyana Goldberg
15: 2013/12/03: Marco De Vivo (ISS Genoa) - Drug Design
16: 2013/12/05: Marco Punta (Pfam, EBI) - Pfam
17: 2013/12/10: Exercise/Presentations
18: 2013/12/12: no lecture: Dies Academicus / Protein-protein interaction 1
19: 2013/12/17: Andrea Schafferhans: 3D function prediction
20: 2013/12/19: Andrea Schafferhans: Docking
21-24: no lectures - winter break (2013/12/23 - 2014/01/06)
25: 2014/01/07: Protein-protein interaction 1
26: 2014/01/09: Protein-protein interaction 2
27: 2014/01/14: Protein-protein interaction 3 / Protein-DNA/RNA interaction
28: 2014/01/16: Protein-DNA/RNA interaction 2
29: 2014/01/21: SNP effect 1
30: 2014/01/23: SNP effect 2
31: 2014/01/28: SNP effect 3
32: 2014/01/30: wrap-up
33: 2014/02/04: examen
34: 2014/02/06: no lecture

© Burkhard Rost (TUM Munich)