Looking for the unusual in Protein structure/function Bioinformatics

BME280B, 30 Sept 2010

Dietlind L. Gerloff
BME Department

gerloff@soe.ucsc.edu

www.soe.ucsc.edu/~gerloff
Main Expertise and Interests in Favorites...

- **Molecule**: Proteins
- **Bioinformatics**: Structural Bioinformatics; Sequence Analysis
- **Biology**: Protein-Protein Interactions; Molecular Evolution
- **Medicine**: Malaria; Infectious Diseases; Immune Response
- **CS**: Data Sharing (Peer-to-Peer); Data Visualization
The Gerloff group is active in many areas...
www.soe.ucsc.edu/~gerloff/group

**Facilitating Software**

- **Data Visualisation/User Interface Design**
  www.malariagenomeexplorer.org
  (UCSC: Herbert Lee)

**Data Sharing between Laboratories**

- **Peer-to-Peer Infrastructure “OpenKnowledge”**
  www.openk.org
  (UCSC: Jonathan Magasin)
Today: Protein-Protein Interaction Challenges + Structural Bioinformatics

**Applied Bioinformatics**

• **CDK/cyclin complex prediction:**
  Partner prediction between paralogous sets

• **Tailored use of structure prediction to get a clue:**
  Molecular models of malarial surface proteins
  HEAT/Armadillo repeat regions in disease genes

**Method Development**

• **1-D Electrostatic Surface Profiles:**
  Site prediction on paralogous domains

(UCSC: Thomas Juettemann, Marcos Woehrmann, Felicia Kemp)
Comparative (Homology) Modelling in a Nutshell

- Template sequences **detectably similar** to targets
- Target-template **alignment** vital for model quality
- Generally amenable for **automation/high-throughput**
- Equivalent to low-med resolution X-ray structure (at best)
A new resource identifying potentially modellable PPI in functional genomics data

BISC: Binary SubComplexes in Proteins Database

Thomas Juettemann


http://bisc.soe.ucsc.edu
1-D Surface Profiles:

A Simplified Representation of Electrostatics on Model Surfaces for Protein Interaction Prediction

New potential PPI sites on Complement Receptor 1 (CR1)

University of Edinburgh:
Shakir Ali
Dinesh Soares
Rupert König
Paul Barlow

UC Santa Cruz:
Marcos Woehrmann
Eric Scott, John Archie

University of Wyoming:
David Liberles
Working with modelled structures

3-D context + Surface properties though approximate

✓ electrostatics
✓ hydrophobicity
but NOT shape!
Large scale modelling as a route to multiple surface comparisons of the CCP-module family

RCAs - Regulators of Complement Activity

CCP modules - Complement Control Protein modules (aka. “sushi” repeats)
Complement Receptor 1 (CR1)

http://www.bionmr.chem.ed.ac.uk/bionmr/public_html/ccp-db.html
Soares et al. PEDI (2005), 18:379-88
Protein-Protein Interactions with CCP-modules

Only few characterised in detail (site + partner) so far

Protein binding sites on CCP-modules can be at different locations (with respect to the conserved structural scaffold)

Models allow visual inspection of electrostatic (GRASP) surfaces of different CCP-modules - but it is impossible to “screen” for similarities and differences in this way.
How would electrostatic surface comparison be helpful?

Conserved function:

• Specific interactions (molecular recognition):
  - conservation pressure better detected by sequence
  - partner proteins differ between homologs
How would electrostatic surface comparison be helpful?

Conserved PPI sites:

• Specific interactions (molecular recognition):
  - conservation pressure better detected by sequence
  - partner proteins differ between homologs

• Unspecific interactions:
  - recurrent observation of a patch within a family (?)

Varying/New PPI sites:

• In some cases, new/different interaction sites may exhibit "excessive" change in surface charge distribution, due to adaptation to the binding partner -

  look for unusually different sites
1-D Electrostatic Profiles

= (context-dependent) charge of solvent-accessible atoms, apportioned per amino acid

**PDB2PQR**
- add hydrogens
- compute charge

**NACCESS**
- > 0.15 accessible

1-D Profile: summed per aa

---

1: Dolinsky et al. (2004), NAR 32:W665-7
2: Hubbard & Thornton (1993)
Order: 3, 61, 1, 6, 9, 10, 18, 18............

Structural Neighbour-Enriched Orders:
Generate structural neighbour-enriched orders (no knowledge of location of binding site):

- Tabulate matrix of spatially adjacent residues (Cβ-Cβ < 8Å)
- Path-walking algorithm favouring 3-D neighbors nearby
- Compute 1000 different combinations of residue orders

**Order1:**
1 27 29 32 31 66 67 68 73 74 4 7 8 18 21 20 33 63 65 75 78 79 82 11 10 9 6 28 23 37 35 19 36 34 59 58 80 81 84 85 49 12 40 41 38 53 52 51 50 86 42

**Order2:**
8 9 7 6 4 28 23 35 18 37 12 10 11 82 79 78 75 74 73 67 68 1 31 20 19 38 40 49 81 80 63 65 66 27 29 32 33 21 36 53 51 50 42 41 86 85 84 52 58 59 34

**Order3:**
78 75 65 63 34 59 58 79 80 82 81 11 9 7 74 73 68 67 66 31 32 33 35 20 19 18 38 53 52 51 50 84 85 86 42 41 40 49 10 8 37 6 28 4 27 1 29 23 21 36 12

Random orders: 85% of windows with 0 or 1 (of 6) pairs adjacent

Structure-based random orders: 73% with 2, 3 or 4 (of 6) pairs adjacent
1-D Electrostatic Profiles

= (context-dependent) charge of solvent-accessible atoms, apportioned per amino acid

For each order:
- compare pair-wise
- sliding window
- “distance” value

Distance matrix (by module)
1000 orders, each 1 tree/clustering
consensus tree

Example: CR1~16, CR1~02 (in one neighbor-enriched order)
Sequence Tree:

“Surface Tree”:

(Ali et al., manuscript in prep.)
(Ali et al., manuscript in prep.)
Our Main Problem: Hardly any information to validate

Able to “screen GRASP cartoons” - but are these binding sites?

**Corroboration / overlay with other clues:**

- Local nature of the signal (surface patch)
- Adaptive Evolution?

**$K_{a}/K_{s}$ or ($K_{a} - K_{s}$) analysis of coding sequences:**

- $K_{a}$: non-synonymous mutation rate
- $K_{s}$: synonymous mutation rate

$\rightarrow K_{a} > K_{s}$ gene is under positive selection
Tertiary Windowing for $K_a/K_s$
(David Liberles, University of Wyoming)

- Local (3-D structural) $K_a$ vs $K_s$ calculations
- Attribute to individual lineages in the evolutionary tree

For example:

**CR1 orthologous genes from: human, chimp, rhesus, baboon**

Tertiary Windowing for $K_a/K_s$
(David Liberles, University of Wyoming)

$K_a > K_s$
~ neutral
too few contacts
Tertiary Windowing for $K_a/K_s$
(David Liberles, University of Wyoming)

$K_a > K_s$ ~ neutral too few contacts
Points of promise...

- The new method can be used as a tool for automatically screening GRASP pictures.

- The sites on CR1~02 and CR1~24 are corroborated by positive selection signals of the gene ($K_a/K_s$ analysis).

- CR1~25 (near site on CR1~24) has been linked to malaria RBC aggregation.
Some protein structures pose particular difficulties for “standard” methods!

A Protocol for Sensitive HEAT/ARM repeat detection
Fred Kippert

Internally Repetitive Folds
Internally Repetitive Folds
Internally Repetitive Folds

Parallel $\beta$-Helix (e.g. pectate lyase)

N C
Internally Repetitive Folds

α-α Superhelix folds (e.g. β-catenin, nuclear importins)
For example:
HEAT, Armadillo and Related (HAR)

Armadillo Repeat $\alpha\text{-}\alpha$ Superhelix (PDB:2BCT)
Upon Closer Inspection
Regularities are not as Striking

HEAT: Elongation factor 3 (PDB:2iw3A) repeats 2-8

ARM: Plakophilin (PDB:1xm9A) repeats 2-8
Misassociation of long $\alpha$-helical proteins with HEAT/Armadillo is commonplace!

Significant risk of error propagation and misannotation

• For example: TOR/ATM/ATR

Controls are as vital in Bioinformatics as in the lab!

Common errors:
- wrong $\alpha$-helical repeat type
- misassigned fold
- one-for-all inference
Sensitive Repeat Detection

Reference Data: 89 HEAT and 41 ARM repeats from known structures, 1215 HEAT and 575 ARM sequences

1. Partition Alignment (Identify Potential Repeat Segments)

Multiple Sequence Alignment (refined, possibly through manual editing; full-length)

(repeat until no further matches)

HHpred(Pfam)
Identify the top-scoring repeat segment (HEAT: PF02985 | ARM: PF00514; E-value < 50) and remove it from input alignment

Multiple Sequence Alignment Segments (each encompassing one potential repeat)

Also examine internal and flanking regions (50-80 aa; overlapping)

2. HHpred/COACH

COACH(Established Repeats)
S = SAM-style reverse scores
HEAT:
Class IV: S ≤ 10
Class III: 10 < S ≤ 15
Class II: 15 < S ≤ 20
Class I: 20 < S
ARM:
Class IV: S ≤ 12
Class III: 12 < S ≤ 20
Class II: 20 < S ≤ 28
Class I: 28 < S

HHpred(Pfam) E = E-value
Class IV: E ≥ 50
Class III: 50 > E ≥ 5
Class II: 5 > E ≥ 0.5
Class I: 0.5 > E

Detection Rate:
HEAT: 83%
ARM: 90%

Predict Repeats if Class I-III
i.e. HHpred E-value < 50 and/or COACH score > 10 (HEAT); > 12 (ARM)

The truth about FAT – sequence analysis of the so-called “FAT domain” in PI3K-related protein kinases reveals a TPR-AD two domain region.

Kippert, Schmid & Gerloff, in prep.

Methods: PSI-BLAST; Alignment; COACH (HMM-HMM); HHPred; FR-servers
HEAT/Armadillo and TPR Proteins are both $\alpha$-$\alpha$ superhelical
The C-terminal part of the TPR-region in TOR displays “typical” peptide binding surface properties.
To close…

• Structure recognition and modelling can provide interesting clues to elucidating PPI

• In our experience particularly electrostatic and hydrophobic surface properties deserve further exploitation

• Beware of “giant” HEAT/ARM proteins we found overprediction in many other cases, e.g. Huntingtin

Contact: gerloff@soe.ucsc.edu