

Defining protein domains - computational and experimental approaches

Christoph Meier
Oxford Protein Production Facility

chrism@strubi.ox.ac.uk



Agenda

- **Introduction to the Oxford Protein Production Facility**
- **Computational Methods for construct design**
 - integrated bioinformatics analysis of targets
OPAL interface
 - RONN Disorder prediction
- **Experimental Methods for construct design**
 - Construct optimization using limited proteolysis
 - Surface engineering of proteins using chemical methods
 - A library-based screening experiment for domain construct identification

Oxford Protein Production Facility

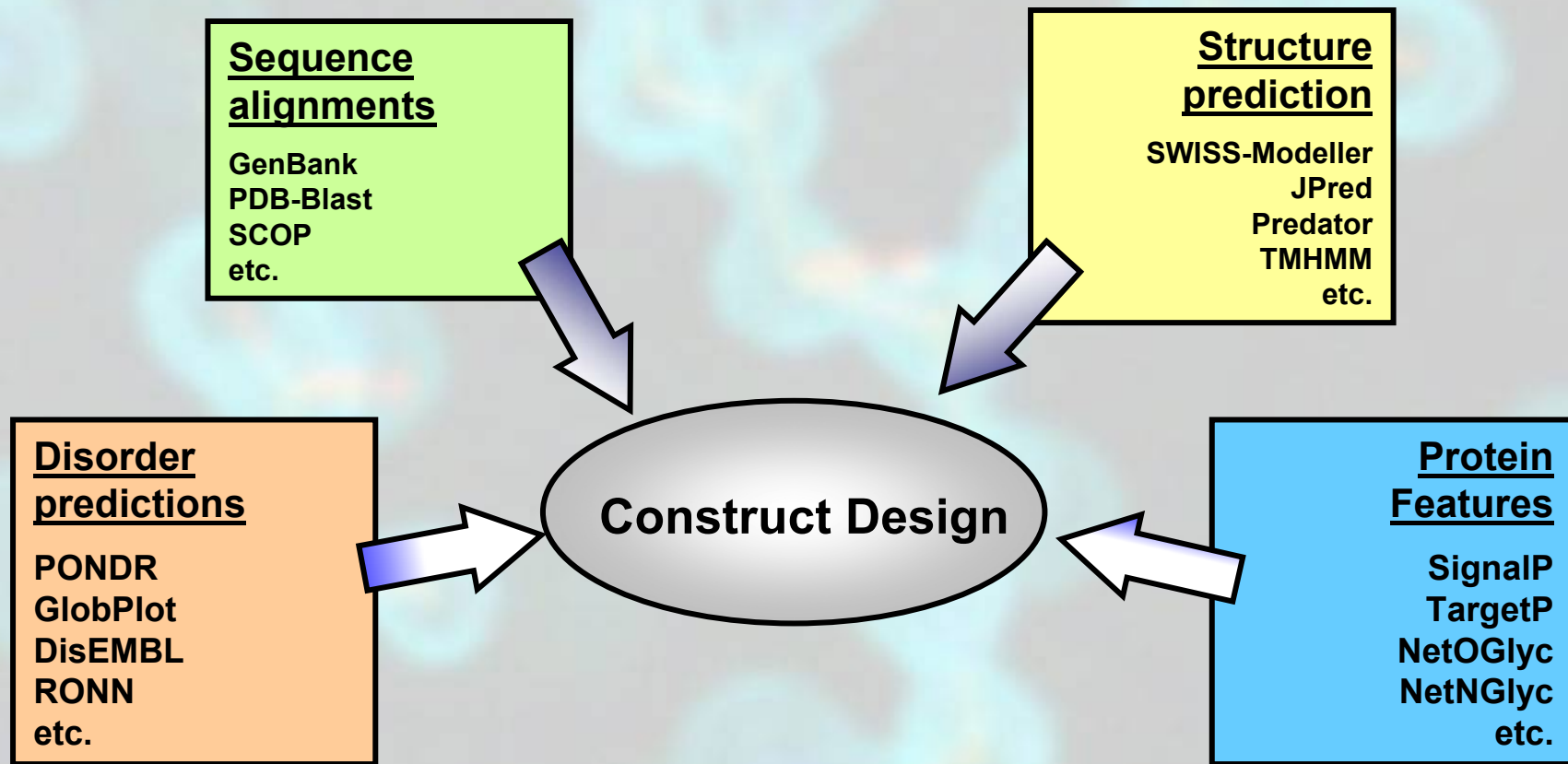
- A high-throughput protein production and crystallisation pipeline.
- Currently a department of the Oxford University.
Future: Diamond Synchrotron (near Oxford)
- Focused on targets of biomedical interest: cancer and immune cell proteomes, virus targets, other human pathogens, e.g. *Neisseria Meningitidis* and *Bacillus Anthracis*.
Targets both intracellular (e.g. Zn-finger transcription factors, SDRs) and extracellular (e.g. secreted and cell surface glycoproteins).
- Pipeline activities: ~1000 targets/year



Computational methods for expression construct design

Construct Design by Bioinformatics

A wide range of computational methods for construct design are available:

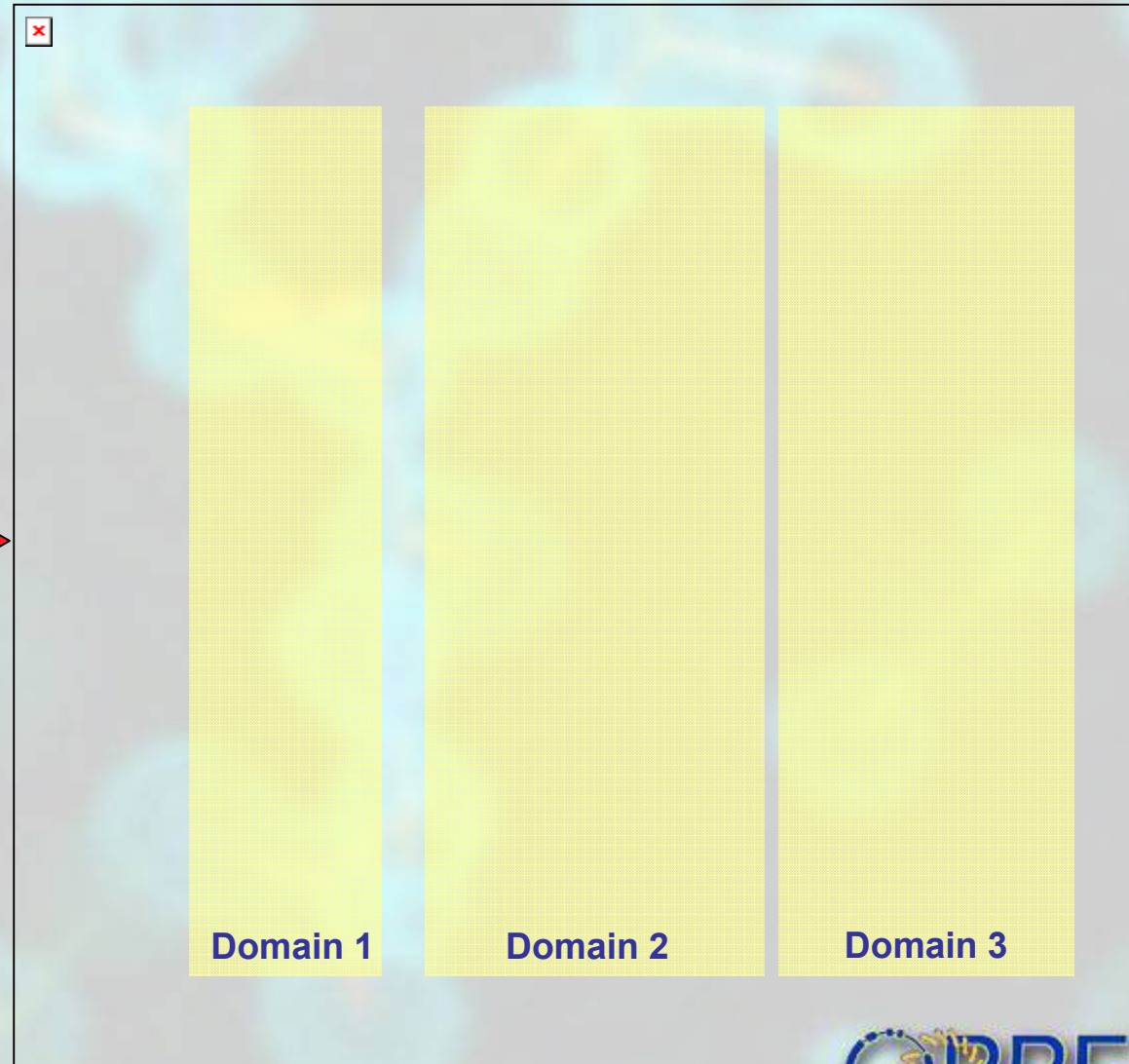


How to integrate different bioinformatics methods?

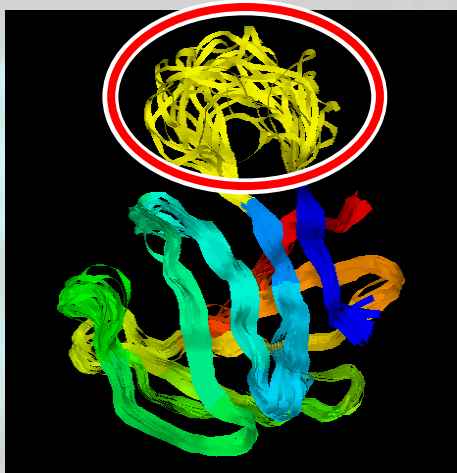
OPAL – Oxford Protein Analysis Linker

<http://www.oppf.ox.ac.uk/bioinformatics.php>

The screenshot shows the OPAL web interface. At the top, there are logos for OPPF, OPAL (Oxford Protein Analysis Linker), and the University of Oxford. A message at the top left states: "If the text on this page is in black, you are using an old version of Netscape. Please upgrade Netscape to use OPAL." Below this is a text input field for a sequence, with a red circle around it and the label "Enter sequence". The text inside the field is a long string of amino acid single-letter codes. Below the sequence field is a text input field for Genbank accession numbers, with a note: "Alternatively (or additionally), enter one or more genbank accession numbers (NOTE: you are NOT recommended to use viral accession numbers!)", and a red circle around it. Below that are radio buttons for organism type: "eukaryote", "gram positive bacteria", and "gram negative bacteria". Then, radio buttons for output type: "full" and "diagram only". A section titled "Choose which programs you want to run on your sequence:" contains a list of analysis tools with checkboxes, including "Codon Usage", "Base content", "GlobPlot", "PSORT", "TargetP", "NetNGlyc", "NetOGlyc", "Protein Calculator", "Predator", "TMHMM", "SignalP", "Blast the Conserved Domain Database", "Blast SCOP", "PDB Blast", "TargetDB Blast", "MDC Clome Blast", "SwissProt Blast", "Genbank nr Blast", "OPTIC database Blast", "Trypsin Cleavage", and "Calculate extinction coefficient". A red arrow points from this section to the right. At the bottom left, there is a "Submit" button and a "Clear Form" button, both circled in red. At the bottom, there is a footer with links to EXPASY, EBI, DIP, BIND, and the Human Protein Reference Database.



Protein Disorder



Intestinal Fatty Acid Binding
Protein (PDB 1A57)

- Many regions of proteins are natively disordered (ie. they fail to self-fold into a specific 3-D structure)
- It is widely accepted that disordered proteins are difficult or impossible to crystallise
- In recent years several disorder prediction algorithms were developed, e.g. DisEMBL, PONDR, GlobPlot, RONN

RONN – Regional Order Neural Network

- Disorder prediction program developed in the OPPF
- Uses a novel neural network (Bio-basis functional network)
- Available free of charge via the internet:

<http://www.strubi.ox.ac.uk/RONN>

or google 'RONN disorder'

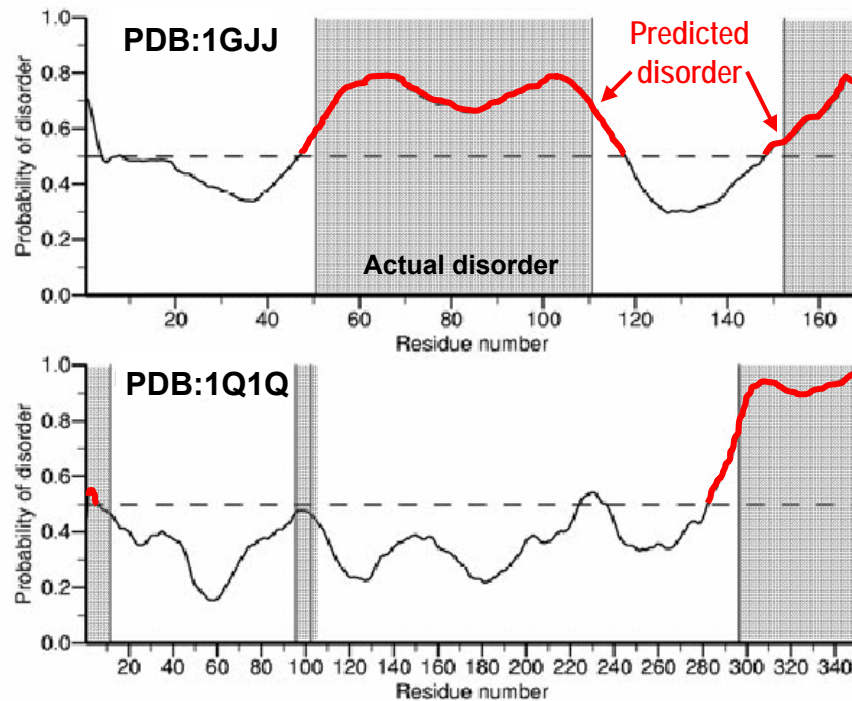


RONN Disorder Prediction

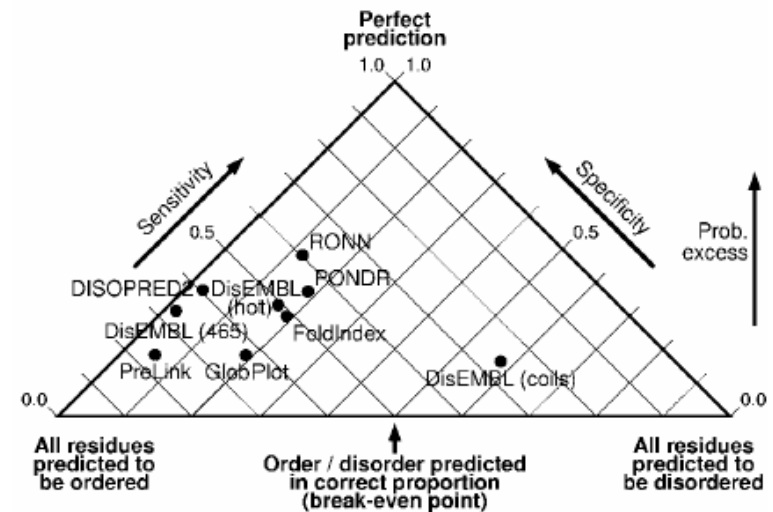
- **How does it work?**

The target sequence are aligned to proteins with known folding properties. The alignment scores are then used to classify each sequence as ordered or disordered using a suitably trained neural network.

- **Examples:**



- **Comparing RONN with other disorder prediction algorithms:**
good sensitivity and specificity

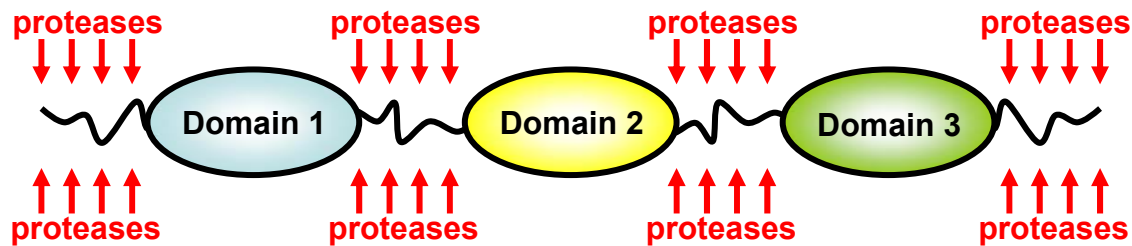


Yang et al. (2005) *Bioinformatics* 21:3369-76

Experimental methods for expression construct design

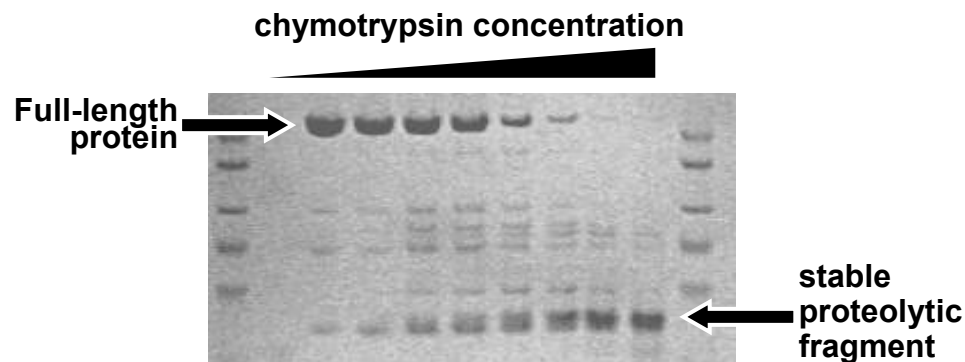
Limited Proteolysis

- compact, globular proteins are good candidates for production of diffraction quality crystals.
- Limited proteolysis: map globular domain(s) of a protein by gently digesting with proteases and identify the fragments (mass spec/ sequencing)



Gao et al. (2005) *Journal of Structural and Functional Genomics* 6:129-134

- Example experiment:



Journal of Structural and Functional Genomics • 6:129-134, 2005
DOI 10.1007/s10000-005-0103-3

High-throughput limited proteolysis/mass spectrometry for protein domain elucidation

Xia Gao, Kevin Bain, Jeffrey R. Bonanno, Michelle Buchanan, Devin Henderson, Don Lorimer, Curtis Marsh, Julie A. Reynolds, J. Michael Sauder, Ken Schwinn, Chai Thai & Stephen K. Burley*
Structural Genomics, Inc., 8550 Resale Street, San Diego, CA 92121, USA *Author for correspondence (e-mail: sburley@strgenomics.com; Fax: +1-619-594-0679)

Received 10 August 2004; accepted in revised form 14 January 2005

Key words: automation, domain definition, limited proteolysis, LFMS, mass spectrometry, protein domain elucidation

Abstract
High-resolution structural information is important for improving our understanding of protein function *in vivo* and *in silico* and providing information to enable drug discovery. The process leading to X-ray structure determination is often time consuming and labor intensive. It requires informed decisions in expression construct design, expression host selection, and strategies for protein production, crystallization and structure determination. Previously published studies have demonstrated that compact globular domains defined by limited proteolysis represent good candidates for production of diffraction quality crystals [1–3]. Integration of mass spectrometry and proteolysis experiments can provide accurate definition of domain boundaries at unprecedented rates. We have conducted a critical evaluation of this approach with 496 target proteins produced by SGJN (Structural Genomics, Inc.) for the New York Structural Genomics Research Consortium (NYSGRC) (<http://www.strgenomics.com>) under the auspices of the National Institute of General Medical Sciences Protein Structure Initiative (<http://www.nigms.nih.gov/psi>). The objectives of this study were to develop parallel automated protocols for proteolytic digestion and data acquisition for multiple proteins, and to carry out a systematic study to compare domain definition via proteolysis with outcomes of crystallization and structure determination attempts. Initial results from this work demonstrate that proteins yielding diffraction quality crystals are typically resistant to proteolysis. Large-scale sub-cloning and subsequent testing of expression, solubility, and crystallizability of proteolytically defined fragments is currently underway.

Introduction
The drive to generate a large number of protein structures of diverse protein families in a cost-effective manner has engendered various parallel-automated solutions to experimental aspects of the problem. Successful structure determination critically depends on making informed decisions throughout the process of construct design, protein expression, purification, crystallization and data collection/processing. Optimal construct selection can greatly facilitate the entire process.

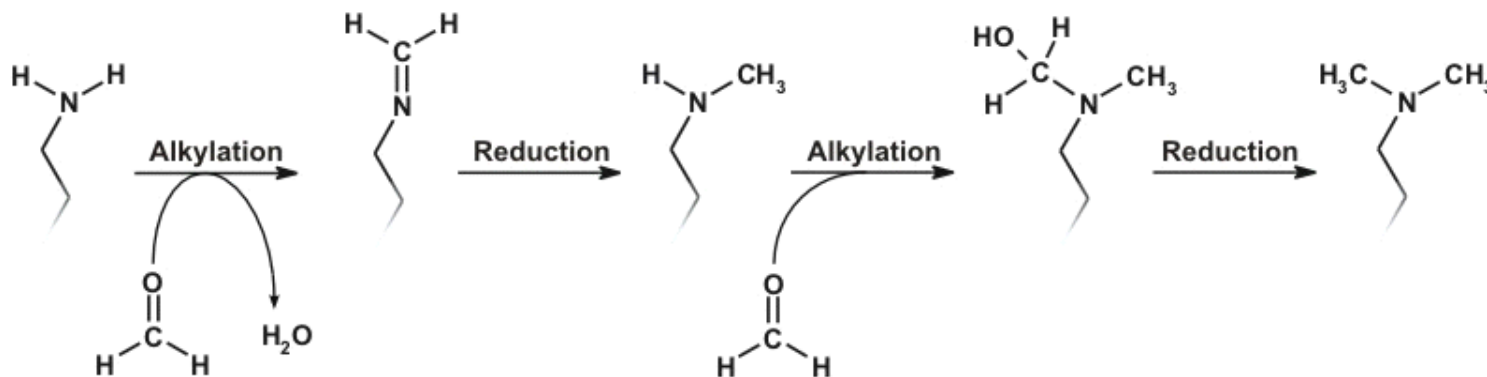
Previous studies [1–3] have shown that globular protein domains can be accurately defined by combining limited proteolysis with mass spectrometry (LFMS). Re-engineered proteins designed on the basis of proteolytic domain definition often exhibit higher solubility and greater propensity to form diffraction-quality crystals.

It is important to conduct LFMS experiments on a large number of proteins and systematically correlated LFMS data with crystallization outcomes because previously published studies mainly



Protein surface engineering as a tool in structural biology

- **Many protein yield no or poorly diffracting crystals**
- Theoretical and experimental studies show:
 - the presence of flexible side-chains (especially lysines) on the surface of proteins reduces crystallizability.
- Therefore
 - promote protein crystallization by chemical modification (reductive methylation) of lysine residues:



Summary

- **Construct design by bioinformatics methods**

a wide range of bioinformatics methods are available (sequence alignments, structure & disorder & feature predictions). We have developed an interface, OPAL that allows these different methods to be integrated.

<http://www.oppf.ox.ac.uk/bioinformatics.php>

- **Experimental Methods**

Expressable proteins which do not crystallize, can be optimized by

- Limited proteolysis:

Can be useful, but usually only in conjunction with other methods.

- Chemical surface engineering:

In our pilot experiment, has proved highly successful.

Quick & Inexpensive – highly recommended.

In collaboration with other labs, we are implementing library-based methods for domain construct identification.

Acknowledgements

Oxford Protein Production Facility

**Dave Stuart, Ray Owens, Jonathan Grimes, Rene Assenberg
Nick Berrow, Sarah Sainsbury, Karl Harlos, Tom Walter**

**Rebecca Hamer, Lester Carter, Robert Esnouf
(OPAL suite & RONN program)**

**Mohammad Bahar, Nicola Abrescia
(Limited Proteolysis)**

Los Alamos National Laboratory

Geoff Waldo, Tom Terwilliger

SPINE and VIZIER collaborators

