(Placeholder) Automatic selection of kinase expression constructs

TBD[1] and John D. Chodera[1,*]

[1]*Computational Biology Program, Sloan Kettering Institute,*
Memorial Sloan Kettering Cancer Center, New York, NY 10065

(Dated: January 17, 2016)

I. INTRODUCTION

Just an outline for now. We want to do large-scale expression testing across the kinase family. Many kinases have already been expressed in various expression systems and with different construct sequences. However, the exact details of the expression (if made available at all) are often buried in the Supplementary Information sections of journal articles. When attempting to carry out expression on a family or superfamily scale, it is not tractable to trawl through hundreds of articles to find relevant expression data. One source of expression construct data which is programmatically accessible is the Protein Data Bank (PDB). Our method is thus based around searching the PDB for relevant expression constructs. PDB data includes expression system and experimental construct, though as we discovered in our research, the latter suffers from frequent problems with misannotation, necessitating us to develop a method to determine authentic experimental sequences.

II. METHODS

A. Semi-automated selection of kinase construct sequences for E. coli expression

1. *Selection of human protein kinase domain targets*

Human protein kinases were selected by querying the UniProt API for any human protein with a domain containing the string "protein kinase", and which was manually annotated and reviewed (i.e. a Swiss-Prot entry). The query string used was:

```
taxonomy:"Homo sapiens (Human) [9606]" AND
domain:"protein kinase" AND reviewed:yes
```

Data was returned by the UniProt API in XML format and contained protein sequences and relevant PDB structures, along with many other types of genomic and functional information. To select active protein kinase domains, the UniProt domain annotations were searched using the regular expression `^Protein kinase(?!; truncated)(?!; inactive)`, which excludes certain domains annotated "Protein kinase; truncated" and "Protein kinase; inactive". Sequences for the selected domains were then stored. The sequences were derived from the canonical isoform as determined by UniProt.

2. *Matching target sequences with relevant PDB constructs*

Each target kinase gene was matched with the same gene in any other species where present, and UniProt data was downloaded for those genes also. The UniProt data included a list of PDB structures which contain the protein, as well as their sequence spans in the coordinates of the UniProt canonical isoform. This information was used to filter out PDB structures which did not include the protein kinase domain - structures were kept if they included the protein kinase domain sequence less 30 residues at each end. PDB coordinate files were then downloaded for each PDB entry. The coordinate files contain various metadata, including an `EXPRESSION_SYSTEM` annotation, which was used to filter PDB entries to keep only those which include the phrase "ESCHERICHIA COLI". The majority of PDB entries returned had an `EXPRESSION_SYSTEM` tag of "ESCHERICHIA COLI", while a small number had "ESCHERICHIA COLI BL21" or "ESCHERICHIA COLI BL21(DE3).

The PDB coordinate files also contain SEQRES records, which should contain the protein sequence used in the crystallography or NMR experiment. According to the PDB documentation (`http://deposit.rcsb.org/format-faq-v1.html`), "All residues in the crystal or in solution, including residues not present in the model (i.e., disordered, lacking electron density, cloning artifacts, HIS tags) are included in the SEQRES records." However, we found that these records are very often misannotated, instead representing only the crystallographically resolved residues. Since expression levels can be greatly affected by insertions or deletions of only one or a few residues at either terminus [DLP: ?CITE, or reference our 96-construct Abl1 expression panel], it is important to know the full experimental sequence, and we thus needed a way to measure the authenticity of a given SEQRES record. We developed a crude measure by hypothesizing that a) most crystal structures would be likely to have at least one or a few unresolved residues at one or both termini, and b) the presence of an expression tag (which is typically not crystallographically resolved) would indicate an authentic SEQRES record. To achieve this, unresolved residues were first defined by comparing the SEQRES sequence to the resolved sequence, using the SIFTS service (CITE) to determine which residues were not present in the canonical isoform sequence. Then regular expression pattern matching was used to detect common expression tags at the N- or C-termini. Sequences with a detected expression tag were given a score of 2, while those with any unresolved sequence at the termini were given a score of 1, and the remainder were given a score of 0.

[*] Corresponding author; john.chodera@choderalab.org

This data was not used to filter out PDB structures at this stage, but was stored to allow for subsequent selection of PDB constructs based on likely authenticity. Also stored for each PDB sequence was the number of residues extraneous to the target kinase domain, and the number of residue conflicts with the UniProt canonical isoform within that domain span.

3. *Plasmid libraries*

As a source of kinase DNA sequences, we purchased three kinase plasmid libraries: the addgene Human Kinase ORF kit , a kinase library from the Structural Genomics Consortium (SGC), Oxford (http://www.thesgc.org), and a kinase library from the PlasmID Repository maintained by the Dana-Farber/Harvard Cancer Center. The aim was to subclone the chosen sequence constructs from these plasmids, though we did not use the same vectors. Annotated data for the kinases in each library was used to match them against the human protein kinases selected for this project. A Python script was written which translated the plasmid ORFs into protein sequences, and aligned them against the target kinase domain sequences from UniProt. Also calculated were the number of extraneous protein residues in the ORF, relative to the target kinase domain sequence, and the number of residue conflicts.

4. *Selection of sequence constructs for expression*

Of the kinase domain targets selected from UniProt, we filtered out those with no matching plasmids from our available plasmid libraries and/or no suitable PDB construct sequences. For this purpose, a suitable PDB construct sequence was defined as any with an authenticity score > 0, i.e. those derived from SEQRES records with no residues outside the span of the resolved structure. Plasmid sequences and PDB constructs were aligned against each target domain sequence, and various approaches were then considered for selecting a) the sequence construct to use for each target, and b) the plasmid to subclone it from. Candidate sequence constructs were drawn from two sources - PDB constructs and the SGC plasmid library. The latter sequences were included because the SGC plasmid library was the only one of the three libraries which had been successfully tested for E. coli expression.

For most of the kinase domain targets, multiple candidate sequence constructs were available. To select the most appropriate sequence construct, we sorted them first by authenticity score (i.e. those with detected expression tags were ranked above those with any other sequence extraneous to the domain span; while those with no extraneous sequence had already been filtered out), then by the number of conflicts relative to the UniProt domain sequence, then by the number of residues extraneous to the UniProt domain sequence span. The top-ranked construct was then chosen. In cases where multiple plasmids were available, these were sorted first by the number of conflicts relative to the UniProt domain sequence, then by the number of residues extraneous to the UniProt domain sequence span, and the top-ranked plasmid was chosen.

This process resulted in a set of 96 kinase domain constructs, which (by serendipity) matched the 96-well plate format we planned to use for parallel expression testing. We therefore selected these construct sequences for expression testing.

A sortable table of results can be viewed at http://choderalab.github.io/kinome-data/kinase_constructs-addgene_hip_sgc.html.

TODO maybe include a figure to help illustrate the above (but may be too complicated):

5. *Other notes*

While much of this process was performed programmatically using Python, many steps required manual supervision and intervention. We hope eventually to develop a fully automated software package for the selection of expression construct sequences for a given protein family, but this was not possible within the scope of this article.

B. Expression testing

TODO For each target, the selected construct sequence was subcloned from the selected DNA plasmid. Expression testing performed by QB3 MacroLab.