Protein spaces

Why?
- We like to compare objects – prediction of properties
- groups of proteins – do they exist?

Spaces – do they exist?
- what is protein space?
- who cares?
A space of protein structures?
who talks about spaces?

Here
- sequence space (proteins)
- structure space (proteins)

Others
- small molecule space – drug space
- tree space
- the set of solutions to a combinatorial problem (not really spaces)
  - how many paths does the travelling salesman problem offer?

What does a space mean to me?
- usually a classic vector space / rarely a discrete space
The questions

I want spaces that are
• objective
• reproducible
• accurate (you cannot have everything)

Proteins
• sequence space (discrete)
• structure space (continuous)
  • sequence and other spaces - continuous
Spaces

Conventional spaces

• 1D ($x$), 2D ($x, y$), 3D ($x, y, z$)
  • 4D ($x, y, z, w$), ...

• let us estimate how big a space or problem is – examples

Sequence alignments – picking penalties
1. gap opening
2. gap widening

The optimal parameters are a point in a 2D space (one point)
Discrete spaces

Discrete space

- how many variables do I have? \((a, b, c, \ldots)\)?
- how many values can each variable have?
  - \(a\) 3 values, \(b\) 4 values, \(c\) 5
  - number of points in space = \(3 \times 4 \times 5\)
Representing a Sequence

Protein sequence and structural coordinates

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>...</th>
<th>$N_{res}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1.2</td>
<td>2.3</td>
<td>...</td>
<td>10.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>2.4</td>
<td>3.5</td>
<td>...</td>
<td>11.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>1.7</td>
<td>2.9</td>
<td>...</td>
<td>15.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>seq</td>
<td>W</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>...</td>
<td>D</td>
</tr>
</tbody>
</table>

A protein is a set of 3D points

A protein is a set of 4D points / descriptors if we add sequence
- 4th dimension is not continuous
- This is NOT sequence space
The sequence points

Usually, a protein is a set of points

I want one point = one protein

Consider proteins of length $N_{\text{res}}$
- look at the first few (3) points
Finding a Sequence in This Space

Real diagram is a box of $N_{res}$ dimensions
• this one 3 dimensions

• looking for sequences...
Families in Sequence Space

Similar sequences are near each other

How realistic?
- only works for \( N_{seq1} = N_{seq2} \)

Conceptual or practical
- important for discussions about protein families (conceptual)
- would you use it directly? maybe with multiple sequence alignments

What is really ugly?
- there is no natural ordering on axes

Summary
- we have a discrete space in which every protein is a point
General continuous spaces

My sequence space
• conceptually useful / practically less so

A generally useful approach
• 2 points fit in 1D (or less)
• 3 points fit in 2D (or less)
• ..
• $N$ points can always fit into $N-1$ dimensions (maybe less)

• my diagrams are usually 2D
Some protein spaces

Do I have a measure of similarity? Many

Sequence-based
- % sequence similarity
- alignment scores
- $k$-mer similarity, ..

Structure-based
- superimpose and look at geometry
- count similarities in secondary structure elements

General rule
- If I can define similarities there is an implied space
How big?

Sequence space? (discrete)
• $20 \times 20 \times 20 \ldots = 20^N$

Conformational space – how to argue
• for each residue, there are at least 2 major groups (really more)
• maybe chop plot into 3 or 5 pieces
  • say there are $c$ conformational possibilities
  • $c^N$ for some $c$
• these spaces grow exponentially in the size of the protein
How general

• You can usually invent a space

• High dimensional spaces are not much fun (directly)
  • what do you do with 7-dimensional coordinates?

What does one normally do?
• reduce to fewer dimensions – find the best 2 or 3-dimensional representation of the data
  • distance geometry (and others) OR

• work with distances - coordinates are just something to think about
Should we expect a hierarchy?

- 7 lowest level clusters
- 3 higher level clusters
- evolutionary argument..
Do we expect protein families?

- No real answer
- We have an idea of spaces – sequence or structure based
- How are proteins distributed?

• Should you expect clusters?
Evolution and phylogeny

Shape / density of tree of life

- ursuppe
- time
- 2015

clear families

no families
• People like to classify proteins

• Should one expect a hierarchy? Maybe

• Should one expect clusters? Yes

• what are the distance measures between proteins?