Why Infinite Dimensional Topological Groups May Work For Genetics

Boris Khots
bkhots@cccgglobal.com
Dmitriy Khots
dkhots@math.uiowa.edu
A significant role in mathematical modeling and algorithms for applications to processing of Genetic data (for example, Gene expression data) may play infinite-dimensional P-spaces and connected with them infinite-dimensional P-groups and P-algebras (B.S. Khots, Groups of local analytical homeomorphisms of line and P-groups, Russian Math Surveys, v.XXXIII, 3 (201), Moscow, London, 1978, 189-190). Investigation of the topological-algebraic properties of P-spaces, P-groups and P-algebras is connected with the solution of the infinite-dimensional fifth David Hilbert problem.

In Genetic data processing the utilization of the topological-algebraic properties of P-spaces, P-groups and P-algebras may permit to find "Gene functionality". We applied these methods to Yeast Rosetta and Lee-Hood Gene expression data, leukemia ALL-AML Gene expression data and found the sets of Gene-Gene dependencies, Gene-Trait dependencies. In particular, accuracy of leukemia diagnosis is 0.97. On the other hand, Genetics requires a solution of new mathematical problems. For example, what are the topology-algebraic properties of a P-group (subgroups, normal subgroups, normal serieses, P-algebras, subalgebras, ideals, etc) that is finitely generated by local homeomorphisms of some manifold onto itself?
Main Idea

- Embedding gene expression data into a known abstract mathematical object
- Viewing the embedded data as a sub-object
- Projecting properties of the object onto the embedded data
- Using the projected properties to analyze the data
Embedding

• The methods of embedding depend on the chosen mathematical object

• The following are examples of embedding gene expression data into a group of homeomorphisms of \([0, 1]\) and, more generally \([a, b]\), onto itself
Embedding – [0,1] Example

- Consider the gene expression data plotted Vs genes
- We will use basic transformations to convert the noisy gene data into a graph of an increasing homeomorphism
- The new graph is now viewed as an element of the group $H([0, 1])$
Embedding – [a, b] Example

• On the left side of the above figure typical sets of gene expression data across a series of n samples are graphically represented as log ratios over the range [a, b]
• In the central part, the data is re-formatted as n(n-1)/2 scatter-plots comparing all possible samples
• Each set of data points can then be embedded into H([a, b]) by reordering as depicted on the right hand side
• For data points that have identical coordinates, a correction based on noise characteristics can be made to satisfy criteria for inclusion into H([a, b])
• We can also embed data into various subgroups of H([a, b])
Mathematical Object

- The most useful objects (since a lot is known of their properties) for our purpose are P-groups.
- P-group is a topological group and an infinite-dimensional manifold which is locally a P-space with smooth operation, smoothness being taken w.r.t. the topology of the given P-space.
- P-space is a linear topological space with certain projection operators & topology.
P-group Example

- \( T(M, M) = \{f: M \to M \mid f \text{ homeo, } M \text{ manifold}\} \) is a P-group
- In particular, in the two examples above, \( H([0, 1]) \) and \( H((a, b)) \) are P-groups
P-group Properties

- Existence of fixed points for all $f$ in $T(M, M)$
- Stationary subgroups
- Normal subgroups
- Composition series
- One-parameter subgroups
- Isomorphism classes of subgroups
\[ f_E, f_{E_i}, f_{E_j}, f_{E_k}, \ldots \in G \]

\[ \text{Wn} \ (f_{E_i}, f_{E_j}, f_{E_k}, \ldots) \rightarrow f_E \]

\( \text{Wn} - \text{word in } G, \ n \rightarrow \infty \)
Using Object Properties

• The above figure depicts a sketch of an algorithm that uses topology-algebraic properties of \(H([a, b])\) for identification of gene expression dependencies

• The injective transformations \(\Phi, \Psi, \ldots, \Xi, \Omega\) map the homeomorphism segments \(E, E_i, \ldots, E_j, E_k\) onto the elements \(f_E, f_{E_i}, \ldots, f_{E_j}, f_{E_k}\) in \(H([a, b])\)

• Gene expression dependency exists iff there is a sequence of the group words \(W_n\) formed by \(E_i, \ldots, E_j, E_k\) in \(H([a, b])\), which converges to \(f_E\) when \(n \to \infty\)
Using Object Properties

- Thus we get a functional dependency of the form $f_E = F(f_{Ei}, \ldots, f_{Ej}, f_{Ek})$

\[ \iff f_x = F(f_1, f_2, f_3, f_4, \ldots f_8) \]
Results

- This approach has identified gene expression dependencies from a well-curated yeast gene expression data set available from Rosetta Inpharmatics
- Using microarray platform, the expression of 6317 genes was profiled from 277 experimental samples (yeast strains, in which a single gene was either deleted or overexpressed) and 63 “controls” (replicate cultures of wild-type yeast)
- Prior to analysis, we removed
  - Data for genes that were expressed below sensitivity of detection in a majority of experiments
  - Data from experiments, where later work showed systematic bias as a result of aneuploidy
- We have found:
  - 24, 691 dependencies of the form $f_{G0}=F(\{f_{Gn}\})$ for 5795 yeast genes
  - Number of genes per dependency: 3-9
  - Number of dependencies per gene: 1-12
  - Total number of genes dependent on each $G_0$: 3-85