Supplementary Text

System Structure : Method description and application

Introduction

One of the authors of this paper developed a set-theoretic, distribution-free, computational model for complex systems, called System Structure from first principles [1]. A complex system can be partitioned into "natural groups", without requiring any a-prior information (including the number of groups), based on the information contained in its System Structure. This characteristic distinguishes this approach from methods widely used for the analysis of population structure e.g STRUCTURE [Pritchard et al, Genetics 155,945(2000)] and makes this approach particularly useful for analysis of heterogeneous and diverse populations. This method was therefore selected as one of the methods of analysis of the Indian Genome Variation data.

In this study, System Structure is first validated as an approach for inferring population structure through the analysis of an earlier data set of 52 populations used in the study of genetic structure of human populations by Rosenberg et al. [Science 298,2381-2385 (2002)]. Subsequently, System Structure and System Structure-based group (referred to as group hereafter) identification is applied for uncovering structure of Indian populations.

Method description

System Structure consists of set-theoretic relations, developed under a generalized construct of uncertainty, between constituent elementary system objects. This definition provides the basis for an iterative computational procedure which is used for quantifying, through a local and system evaluation, potentials associated with the elementary system objects and their inter-relations. The term System Structure also equivalently refers to a weighted network representation of a system: The nodes and arcs are respectively symbolic of the elementary system objects and their interrelations, and both are weighted by system measures which satisfy the set-theoretic relations [1].

This method does not assume a restrictive a prior network model or make specific assumptions about the stochastic process underlying the network formation and yet provides a systematic quantification of the underlying uncertainty. Thus, in contrast to Bayesian approaches, this method is less computationally intensive and is also less model-dependent. Furthermore, in contrast to classical methods of statistical inference, this method does not make assumptions about independence of events or require as many experimental replicates/ samples for arriving at a robust inference. This makes it particularly useful as it obviates the need for very large samples required by classical statistical methods to obtain sufficient power.

System Structure thus provides a useful mathematical representation, facilitating both the conceptualization and computational implementation of a number of methodological problems. These include the identification of influential nodes and significant sub-networks (communities/ partitions/ group structures), classification, and time series analysis.

This portfolio of methods has been applied to a number of diverse systems and problems including quantification of spillovers in innovation systems, sleep staging, quantification of drowsiness, electroencaphalogram—based prediction of epileptic seizures and associations studies with hypertension [2 - 9].

A detailed documentation about the theory and algorithm underlying System Structure is also available online thesis (http://physio1.eecs.cwru.edu:8802/~amit/)

Application to Analysis of Genetic Variation:

System Structure has been applied within two frameworks of genetic variation analysis: In the first framework, that of discovering substructures, unlabeled samples and loci are identified as elementary system objects. The samples and loci are connected through a bipartite network with each sample connected to each loci with a weight corresponding to the available sample-loci genetic information (genotype, microsatellite length, bi-allelic information etc.). System Structure, applied to this weighted network, converges with the

corresponding Sample-loci System Structure: A bipartite network between samples and loci with the nodes and arcs are weighted by system measures which quantify the associated potentials. This Sample-loci System Structure provides an informative representation of the genetic structure and is used to compute a Sample-Sample System Structure.

In the second framework, System Structure again assumes a bipartite network with one set of nodes corresponding to pre-specified population labels and with the second set corresponding to the loci and the arc weights as the corresponding genotype/ allele frequency. In this case, population-loci system measures are used to compute a fully connected population-population System Structure with the arcs weighted with population-population system measures.

System Structure Based Community (Group) Identification:

Sample-sample (population-population) network is partitioned into significantly connected, distinct communities of populations (sub-networks) through a nested thresholding approach: the first threshold is on the system measures between populations for identifying significant population set associated with each population and the second threshold is on the overlap between these sets.

Thus, groups identified correspond to significantly connected 'heavy' (in terms of higher arc system measures) network partitions. The optimal grouping is identified through an entropy evaluation of three types of measures: fuzzy, possibility and typicality generated for each sample (population). System Structure based group identification also generates fuzzy and possibility measures between communities which are also evaluated for entropy.

Fuzzy measures quantify a population membership across groups while possibility memberships quantify a population's membership within a group. All samples (populations) belong to all groups with varying fuzzy and **possibility** memberships while being assigned to a group in which they have the maximum possibility membership. In addition some members are identified as core or prototypical of each group and **typicality** measures are computed for these.

We find that fuzzy and possibility memberships are indicative of admixture. For example, in some cases, a sample (population) may have highest fuzzy membership and highest possibility memberships in different groups. Such a sample may be an outlier while being most proximal to a strongly homogeneous group in which it has the highest fuzzy membership and on the other hand more similar to other members in a highly heterogeneous group in which it has the highest possibility membership.

As an example, using data of Rosenberg et al (2002) we have analysed human genetic variation using System Structure method (**Figure 1**). System Structure reveals 7 optimal groups in the dataset compared to K=6 in Rosenberg's study using STRUCTURE. We observe that all samples from America have the highest fuzzy and possibility memberships in the same group (Group 1). Samples from Africa also group together with highest fuzzy and possibility (Group 3). On the other hand, many samples from Central/ South Asia and some from East Asia have highest fuzzy and possibility membership in different groups. Substructures obtained with maximum assignment based on possibility memberships are likely to be more representative than substructures obtained with maximum assignments based on fuzzy membership; Clearly, the fuzzy membership-based rule can assign a sample to the group to which it is most proximal but may increase the heterogeneity of the group. Nonetheless, substructures obtained with fuzzy membership-based assignment are informative when viewed as providing additional information about substructures obtained with possibility memberships.

In our study, as with the results reported with STRUCTURE by Rosenberg et al (2002), both Burusho and Kalash tend to isolate independently in possibility memberships in Group 2 and Group 5 respectively, with some admixtures from neighboring populations.
All samples from Africa tend to group together in possibility memberships. Though there is evidence of higher admixture in membership values there is no change in group assignments. Fuzzy memberships of samples from Europe, Central-South Asia, and Middle-East show evidence of admixture with the African members. Using System Structure, all samples from America tend to group together and separately from the rest

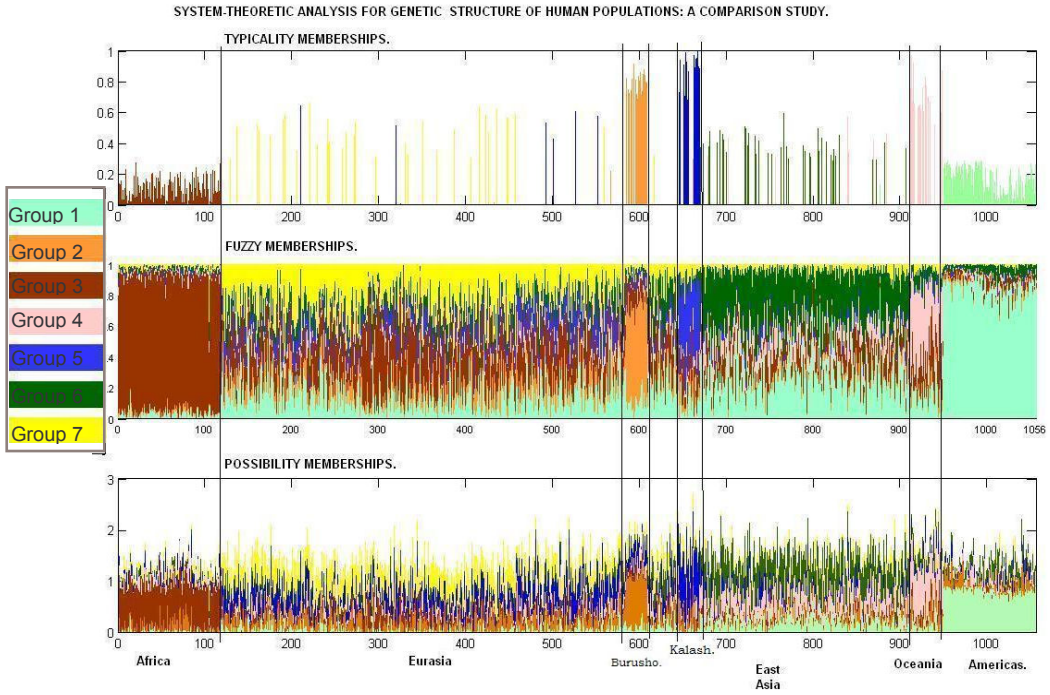
in both fuzzy and possibility. This corroborates with the findings reported in Rosenberg's study using STRUCTURE.

Typicality memberships need to be interpreted carefully as the typicality memberships in a group are relative only to the core members of that group. Core members of a strongly homogeneous group will have lower typicality than some core members of a heterogeneous group. Thus in a highly homogeneous group, a core member is as good as the other core members. On the other hand, in a heterogeneous group, some members are significantly more representative of the group than the rest.

For example, samples from America, which tend to all group together with very little admixture, have a maximum of 0.2 typicality membership (corresponding to a Surui sample from Brazil) and an average typicality membership of 0.1. Of the 121 samples assigned to this group 118 samples belong to the core. On the other hand, a highly heterogeneous group of samples admixed from Europe, Middle-East, and Central—South Asia shows a maximum typical membership of 1 (for a sample from France) and an average typicality membership of 0.56. However, of the 95 samples assigned to this group only 29 are deemed typical.

As already mentioned, entropy-minimizing partitions are obtained based on these three types of memberships: fuzzy, possibility and typicality. Also, as discussed above, these measure different properties; fuzzy measures the variability of a sample (population) across groups, possibility measures the homogeneity within a group and typicality measures the homogeneity within the core members of the groups. Optimal partitions correspond to a concurrent minimization of entropy based on all three measures along with the criterion that the entropy values do not change too much with a change in the second threshold of overlap in associated significant sample sets. This process increases the likelihood that the partitions thus obtained are robust and the measures provide an information-rich basis for inference.

The distinct advantage of System Structure as a model for genetic variations is that the information generated by this approach compares favorably with the information generated by model-based methods as a basis for inference about population substructures without being as computationally intensive. This observation is supported by several genetic variation studies being currently analyzed some of which are reported here. Furthermore, System Structure-based group analysis does not require an a prior specification of the number of groups; instead the group identification process is guided by an entropy evaluation that identifies information maximizing groupings. While System Structure is definitely more computationally intensive than distance-based methods, the measures generated by this systems approach are far more informative than distance; system measures capture both linear and non-linear inter-relations and are not dominated by global patterns.



[1] A. K. Sinha, A Fuzzy Measure-Theoretic Quantum Approximation of an Abstract System. PhD thesis, Case Western Reserve University, January, 2001.

[2] M. S. Fogarty, "Cleveland's emerging economy, a framework for investing in education, science, and technology," tech. rep., Center for Regional Economic Issues, 1998.

[3] M. S. Fogarty, A. Sinha, and A. B. Jaffe, "Advanced technology program and the US innovation system— a methodology for identiying enabling R and D spillover networks with an application to

microelectro-mechanical systems (MEMS) and optical recording,”,” tech. rep., National Bureau of Economic Reviewers, Cambridge MA, USA., 1999.

[4] M. S. Fogarty and A. K. Sinha, “University-industry relationships and regional innovation systems, why older regions can’t generalize from route 128 and silicon valley,” in Industrializing Knowledge: University—Industry Linkages in Japan and the
United States, (L. M. Branscomb, F. Kodama, and R. Florida, eds.), Cambridge, MA. USA: MIT Press, 1999.

[5] M. S. Fogarty, A. K. Sinha, and A. Jaffe, “Sustaining the “new economy” california as a source of new technology,” tech. rep., Public Policy Institute of California, 1999.

[6] A. K. Sinha, W. J. Richoux, and K. A. Loparo, “A system theoretic state description for temporal transitions in electroencephalogram data of severe epileptic patients,” (Atlantis, Paradise Island, Bahamas), 43rd IEEE Conference on Decision and Control, December 2004.

[7] A. K. Sinha, K. A. Loparo, S. Redline, and et al., “Temporal features of sleep predict hypertension: Application of a novel analysis of sleep stage data from the sleep heart health study,” (Philadelphia, Pensylvania), American Professional Sleep Society Annual Meeting, June 2004.

[8] A. K. Sinha, K. A. Loparo, and W. J. Richoux, “A new system-theoretic classifier for detection and prediction of epileptic seizures,” (San Francisco, California), 26th Annual International Conference IEEE Engineering in Medicine and Biology Society (EMBS), September 2004.

[9] Fogarty, Michael S., Amit K. Sinha, and Adam B. Jaffee., ATP and the U.S. Innovation System: A Methodology for Identifying Enabling R&D Spillover Networks, NIST GCR 06-895, October 2006.

[10] A. K. Sinha and K. A. Loparo, A Mathematical and Computational Model for the Analysis of Complex Systems with Applications in Biological and Engineering Systems. Springer-Verlag, forthcoming, 2007.