magnet even on such massive amounts of sequence fragments. In summary, the combination of a sound statistical machinery with a highly engineered algorithm allows for implementation of a reversed discovery workflow.

As a result, the Genome of the Netherlands project is the first of its kind to exhaustively report on the corresponding class of genetic variants, previously termed "twilight zone deletions and insertions", but which now enjoy somewhat more daylight.

In future work, we are also planning to eliminate this blind spot in somatic variant discovery, which will likely reveal large amounts of so far undetected cancer-causing genetic variants, and will hopefully shed considerable light on cancer biology as well.

Links:
http://homepages.cwi.nl/~as
http://www.nlgenome.nl

References:

[1] T. Marschall, I. Hajirasouliha, A. Schönhuth: "MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels", Bioinformatics 29(24):3143-3150, 2013.
[2]The Genome of the Netherlands Consortium: "Whole-genome sequence variation, population structure and demographic history of the Dutch population", Nature Genetics 46(8):818-825, 2014.
[3] W. Kloosterman, et al.: "Characteristics of de novo structural changes in the human genome", Genome Research 25:792-801, 2015.

Please contact:

Alexander Schönhuth
CWI, The Netherlands
E-mail: A.Schoenhuth@cwi.nl

Tobias Marschall was a postdoc at CWI from 2011-2014. Since 2014, he holds an appointment as assistant professor at the Center for Bioinformatics at Saarland University and the Max Planck Institute for Informatics in Saarbrücken, Germany

Computational Estimation of Chromosome Structure

by Claudia Caudai and Emanuele Salerno

Within the framework of the national Flagship Project InterOmics, researchers at ISTI-CNR are developing algorithms to reconstruct the chromosome structure from "chromosome conformation capture" data. One algorithm being tested has already produced interesting results. Unlike most popular techniques, it does not derive a classical distance-to-geometry problem from the original contact data, and applies an efficient multiresolution approach to the genome under study.

High-throughput DNA sequencing has enabled a number of recent techniques (Chromosome Conformation Capture and similar) by which the entire genome of a homogeneous population of cells can be split into high-resolution fragments, and the number of times any fragment is found in contact with any other fragment can be counted. In human cells, the 46 chromosomes contain about three billion base pairs (3 Gbp), for a total length of about 2 m, fitting in a nucleus with a radius of 5 to 10 microns. As a typical size for the individual DNA fragments is 4 kbp, up to about 750,000 fragments can be produced from the entire human genome. This means that there are more than 280 billion possible fragment pairs. Even if the genomic resolution is substantially lowered, the resulting data records are always very large, and need to be treated by extremely efficient, accurate procedures. The computational effort needed is worthwhile, however, as the contact data carry crucial information about the 3D structure of the chromosomes: understanding how DNA is structured spatially is a step towards understanding how DNA works.

In recent years, a number of techniques for 3D reconstruction have been developed, and the results have been variously correlated with the available biological knowledge. A popular strategy to infer a structure from contact frequencies is to transform the number of times any fragment pair is found in contact into the distance between the components of that pair. This can be done using a number of deterministic or probabilistic laws, and is justified intuitively, since two fragments that are often found in contact are likely to be spatially close. Once the distances have been derived, structure estimation can be solved as a distance-to-geometry problem. However, translating contacts into distances does not seem appropriate to us, since a high contact frequency may well mean that the two fragments are close, but the converse is not necessarily true: two fragments that are seldom in contact are not necessarily physically far from each other. Furthermore, we checked the topological consistency of the distance systems obtained from real data, and found that these are often severely incompatible with Euclidean geometry [1].

For these reasons, we chose to avoid a direct contact-to-distance step in our technique. Another problem we had to face when trying to estimate the chromosome structure was the above-mentioned size of the data record, and the related computational burden. The solution we propose exploits the existence of isolated genomic regions (the Topological Association Domains, or TADs) characterized internally by highly interacting fragments, and by relatively poor interactions with any

other segment of the genome. This allows us to isolate each TAD and reconstruct its structure from the relevant data set, independently of the rest of the genome, then lower the resolution, considering each TAD as a single chain element, and then take the weaker interactions between TAD pairs into account, in a sort of recursive, multiresolution approach.

The result is an algorithm (CHROMSTRUCT [2]) characterized by:

- A new modified-bead-chain model of the chromosomes;
- A set of geometrical constraints producing solutions with consistent shapes and sizes;
- A likelihood function that does not contain target distances derived from the contact frequencies – in the present version, this likelihood is sampled by a Monte Carlo strategy to estimate a number of feasible structures for each TAD;
- A recursive framework to associate the structure of each reconstructed TAD with the shape and the size of a single bead in a lower-resolution chain, whose structure, in turn, is estimated on the basis of an appropriately binned data set;
- A recursive framework to build the final structure from the partial results at the different levels of genomic resolution.

So far, we have tested our algorithm on part of the human genome (29.2 Mbp from chromosome 1, at 100 kbp resolution, see Figure 1). The geometrical features of many of our results correlate positively with known functional features of the cells considered in our tests. To conclude our research, and to be able to assess our results against more detailed biological properties, we still need to remove the experimental biases from the raw data, and then try our strategy on larger parts of (or an entire) genome.

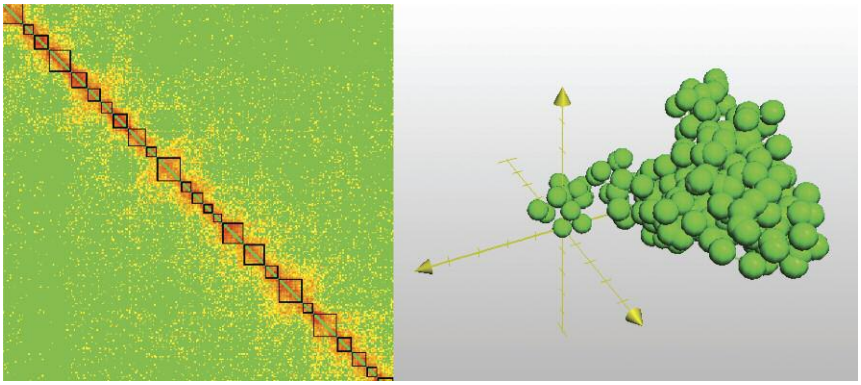

Figure 1: Left: Contact frequency matrix for a segment of the long arm of chromosome 1 (q from 150.28 Mbp to 179.44 Mbp) from human lymphoblastoid cells GM06990, in logarithmic colour scale. Data from [3]; genomic resolution 100 kbp. The highlighted diagonal blocks define our maximum-resolution TADs. Right: one of our reconstructed structures, consisting of a chain with 292 beads.

Link:

InterOmics Flagship Project:
http://www.interomics.eu/web/guest/home

References:

[1] C. Caudai, et al.: "A statistical approach to infer 3D chromatin structure", in V. Zazzu et al. (Eds.), Mathematical Models in Biology, Springer-Verlag, to appear, DOI: 10.1007/978-3-319-23497-7_12.

[2] C. Caudai, et al.: "Inferring 3D chromatin structure using a multiscale approach based on quaternions", BMC Bioinformatics, Vol. 16, 2015, p. 234-244, DOI: 10.1186/s12859-015-0667-0

[3] E. Lieberman-Aiden, et al., "Comprehensive mapping of long-range interactions reveals folding principles of the human genome", Science, Vol. 326, 2009; pp. 289-293, DOI: 10.1126/science.1181369.

Please contact:

Claudia Caudai, Emanuele Salerno
ISTI-CNR, Italy
E-mail: claudia.caudai@isti.cnr.it, emanuele.salerno@isti.cnr.it

Modelling Approaches to Inform the Control and Management of Invasive Seaweeds

by James T. Murphy, Mark Johnson and Frédérique Viard

Invasive non-native plant and animal species are one of the greatest threats to biodiversity on a global scale. In this collaborative European project, we use a computer modelling approach (in association with field studies, ecological experiments and molecular work) to study the impact of an important invasive seaweed species (Undaria pinnatifida) on native biodiversity in European coastal waters under variable climatic conditions.

The introduction of non-native species can transform habitats and disrupt ecosystems resulting in serious environmental and economic consequences. Non-native seaweeds represent one of the largest groups of marine invasive organisms in Europe. However, often the fundamental processes that affect their population dynamics and invasion success are poorly understood making it difficult to develop optimal management strategies at both a local and international scale.

The Asian kelp species *Undaria pinnatifida* (Wakame) has been nominated as one of the world's 100 worst invasive species according to the Global Invasive Species Database [1]. This species is the focus of a collaborative European research project supported by an Irish-Research Council-Marie Curie Actions co-funded Elevate international career development fellowship (2013-