Comparing Genomes

Bud Mishra[a]
a. Professor of Computer Science & Mathematics (Courant Institute, NYU)
Professor (Cold Spring Harbor Laboratory) [1]

Can it then be that there is... something of use for unraveling the universe to be learned from the philosophy of computer design?
– J.A. Wheeler, *Int. J. Theor. Phys.*, **21** 557 (1982).

1 Introduction

As new approaches continue to be developed for the purpose of using biological material to solve difficult computational problems, several fundamental questions are beginning to be asked: *Are these techniques practical? What are the key applications? Do these techniques scale to larger problems? Ultimately, is this a productive endeavor? Do they provide us more than few elegant theoretical insights into the nature of computation?*

Before answering these questions, it may be fruitful to examine the following quote from Richard Feynmann, as it reflects on similar questions in the context of quantum-mechanical computers:

> "*The discovery of computers and the thinking about computers has turned out to be extremely useful in many branches of human reasoning. For instance, we never really understood how lousy our understanding of language was, the theory of grammar and all that stuff, until we tried to make a computer which would be able to understand language. We tried to learn a great deal about psychology by trying to understand how computers work. There are interesting philosophical questions about reasoning, and relationship, observation, and measurement and so on, which computers have stimulated us to think about anew, with new types of thinking. And all I was doing was hoping that computer-type of thinking would give us some new ideas, if any are really needed.*"
>
> – R. Feynmann, "Simulating Physics with Computers,"
> *Int. J. Theor. Phys.*, **21** pp 486, (1982).

[1] *Work reported here has been supported by the following Research Grants: "High-Density Gene Copy Number Microarrays," National Institutes of Health; "Genomics via MicroArrays," NYU University Research Challenge Fund; "Bioinformatics Prototyping Language for Mapping, Sequence Assembly and Data Analysis," Department of Energy; "Faculty Development Program for Bioinformatics and Genomics," New York State Office of Science, Technology, & Academic Research (NYSTAR); "Algorithmic Tools and Computational Frameworks for Cell Informatics," DARPA and "Algorithmic and Mathematical Approaches in Cell Informatics," HHMI Biomedical Support Research Grant.*

In a similar vein, we will put forth our arguments that, while the approaches of biocomputing is based on several classical biotechnological tools such as *restriction activities, hybridization, ligation, PCR and cloning* – just to name a few –, ultimately the reasoning and design style emerging in the field of biocomputing will lead to more sophisticated, robust and high through-put biotechnology. Or at least, that is one aspect of the approach that should not be overlooked.

In order to develop these ideas, we will look at just one application, involving comparison of two related genomes. This problem has many applications in cancer research and was originally developed in collaboration with my colleague Mike Wigler and his laboratory at Cold Spring Harbor. But the description here will simply focus on the computational aspect of the problem and uses some ideas from a recent paper by Casey, Mishra and Wigler [4].

2 Comparing Genomes

The motivation for the problem, we describe below, comes from our efforts to understand the genetic basis of cancer. Roughly, in order to deduce what makes a cell go into uncontrolled growth, we need to focus on the genes involved in a cell making important decisions about growth, growth arrest and apoptosis (cell death). The genes involved in these processes fall into two categories: about one hundred oncogenes and about a thousand tumor suppressor genes.

The way a healthy cell deviates from its normal function to initiate tumor formation is caused by various changes to the genome: *amplifications*, *deletions*, *translocations* and *point mutations*. Both amplification and deletion result in fluctuations of the copy-number of the genes: either increase in case of amplification or decrease in case of deletion. Thus detection of regions of amplification can lead us to the locations of oncogenes and regions of deletion, to tumor suppressor genes. Thus the differences between the genomes from healthy tissue versus cancer tissue tell us a lot about where the oncogenes and tumor suppressor genes may be located.

However, comparing two genomes rapidly appears to be an elusive goal. Recently, Hanahan and Weinberg stated pessimistically:

> "*At present, description of a recently diagnosed tumor in terms of its underlying genetic lesions remains a distant prospect. Nonetheless, we look ahead 10 or 20 years to the time when the diagnosis of all somatically acquired lesions present in a tumor cell genome will become a routine procedure.*"
>
> – D. Hanahan and R. Weinberg, Cell, **100**: 57-70, (2000).

Clearly, we cannot simply sequence the genomes completely and compare, as such an approach will not be cost-effective for the foreseeable future. Instead, we focus on a randomized approach, quite common in the field of computer algorithms. We can sample the genome uniformly to create a large number of probes (150,000) located every 20 Kb (expected distance) and each probe, almost surely unique.

These probes are short subsequences of length 200 to 1200 base pairs and come from the regions of genomes that do not share "homologous" sequences somewhere else in the genome. Our approach then reduces to determining the relative locations of these probes in the two genomes: in terms of their relative ordering, in terms of their presence (possibly multiple times) or absence, or just, in terms of the changes to their relative distances within a small chromosomal region.

Thus, if we can create an inexpensive biotechnological method to measure the distances between any two probes, then the focus of our research moves to the algorithmic problem of finding the locations of these probes along the two genomes, or even the simpler problem of determining when the relative locations of a small group of closely clustered probes are perturbed from one genome to another. Of course, the bio-chemical method we will develop will also be exposed to the corrupting effects of many independent error sources and modeling these errors will be a key challenge for us.

The fundamental idea of our algorithm, which localizes the probes along the genome, is based on the simple observation that if one can determine the pair-wise distances among all the probes, then one can place these probes along the genome correctly. If the distances are known accurately, then for any three probes a triangle-equality is satisfied and with the known locations of any two of the three probes, the location of the third probe is uniquely determined. When the pair-wise distance data are inaccurate, then the triangle-equality (and other similarly higher-order constraints) will be violated and the distance data is inconsistent. Thus, the algorithmic question becomes: "*How can the distance-data be minimally perturbed so that they become consistent?*" and such a question can be formulated as an optimization problem for a weighted sum-of-square cost function. Although, in the most pathological context, such problems can be computationally infeasible, we have developed a simple almost-linear-time probabilistic algorithm that works well for a carefully designed experiment (e.g., choosing the expected number of probes per clones and number of hybridization experiments, etc.) [4]

Thus the focus of our research moves to the following key questions: *How do we model the errors in the distance function and how do we design the parameters of the experiments?*

Roughly, a single biochemical hybridization experiment (conducted with a microarray) assigns a discrete value (a "color": B = Blank, R = Red, G = Green and Y = Yellow) to each probe. A sequence of such experiments will assign a "color vector" to each probe and the number of places in which these color vectors differ for any two probes will give us a clue about the distance between these two probes. Thus the distance metric between two probes is derived from a Hamming distance between every pair of color vectors assigned to the probes. As we conduct a succession of these hybridization experiments, the Hamming distance between two probes is incremented by one every time the probes disagree on the outcomes of any hybridization experiment. Thus the probabilistic modeling of the errors in distance simply involves deriving a conditional probability that the two probes will disagree in an experiment given that they are some given distance apart.

3 Tools of the Trade

We start with some biological background, leading to three key biotechnological tools: The tools of our trade. The usual configuration of DNA is in terms of a *double helix* consisting of two *chains* or *strands* coiling around each other with two alternating grooves of slightly different spacing. The "backbone" in each strand is made of alternating big sugar molecules (Deoxyribose residues) and small phosphate molecules.

One of the four bases (the letters in an alphabet $\Sigma = \{A, T, C, G\}$), each one an almost planar nitrogenic organic compound, is connected to the sugar molecule. The bases are: *adenine* (A), *thymine* (T), *cytosine* (C) and *guanine* (G). So, if one reads the sequence of bases then that defines the information encoded by the DNA. Complementary base pairs (A-T, and C-G) in the two strands are connected by hydrogen bonds and the base-pair forms an essentially coplanar "rung" connecting the two strands. This Watson-Crick complementarity is what makes a DNA chemically inert and mechanically stable, and hence, an ideal molecule for various mechanical and computational devices.

However, these DNA molecules can be manipulated with various biochemical tools: Scissors, Glues and Copiers.

- **Scissors**, Restriction Activity: Type II sequence specific restriction endonucleases are enzymes that can "cut" a double-stranded DNA by breaking the phosphodiester bonds on the two DNA strands at specific target sites on the DNA. These target sites or "restriction sites" are determined completely by their base-pair composition—usually, a very short sequence of base-pairs with their lengths varying from 4 to 8. For instance, the restriction enzyme *Hpa* II will cut the DNA anywhere there is an occurrence of the tetranucleotide $CCGG$.

  The type II restriction endonucleases evolved in nature as the bacterial immune system against the viral DNA; bacteria use these enzymes by cleaving (or "restricting" the activity of) invading foreign DNA.

  The enzymes have been extremely useful in biotechnology as biochemical "scissors" and biochemical "markers" as they always cut DNA at the same short specific patterns (of length 4, 6 and 8). In our application, we will use restriction enzymes to cut a genome into small pieces and then only select a subset of these fragments for further use as probes. As a result, the probes generated this way are reproducible, reliable and consistent. Furthermore, parallel representations (probe sets selected from two genomes) preserve gene ratios and hence provide a crucial tool for our application.

- **Glues**, Ligation and Hybridization: In contrast, DNA ligase is a cellular enzyme that can join two strands of DNA molecules by repairing a phosphodiester bonds. We will not make explicit use of DNA ligases in our application here, but it enjoys wide-spread usage as a key biotechnological tool.

Our focus will be on the process of hybridization, which uses hydrogen bonding between two complementary single stranded DNA fragments (or an RNA fragment and a complementary single stranded DNA fragment) to create a double-stranded DNA (or a DNA-RNA complex). In the current application, the primary use of hybridization is in detecting if a short string (e.g., probe) appears as a substring in a longer string (e.g., a clone or subgenomic DNA). In order to achieve this, we can create a DNA fragment encoding the complementary sequence for the probe and conduct an experiment to see if the complementary-probe-sequence hybridizes to a DNA fragment encoding the longer sequence.

The method can be parallelized by spotting on a surface several probe sequences as a matrix of a very large number of spots (several thousand) and hybridizing all the probes with one or more clone-sequences in parallel. If more than one clone-sequences are involved then this approach allows us to verify if a particular probe-sequence belongs to any one of the clone sequences. This technology embodied as microarrays enjoys wide-spread applications in measuring gene-expressions, classifying genes, mapping markers on the genome and in detecting polymorphisms.

- **Copiers**, Cloning and PCR: For our purposes, a clone is a rather large fragment of a DNA that has been pre-selected and kept in a library, and one can make faithful copies of this DNA fragment many many times. The size of a clone can be 1–2 Mb (YAC, Yeast Artificial Chromosomes), 100–200 Kb (BAC, Bacterial Artificial Chromosomes), 20–45 Kb (Cosmids) or 2–20 Kb (lambdas).

  The molecular cloning is an *in vivo* approach involving a living host organism (usually the E. coli bacteria or yeast) which replicates a suitably modified foreign DNA, as if the foreign DNA is one of its own DNA. The modification involves combining a cloning "*vector*" with the foreign DNA to be amplified, the "*insert*," to create a circular recombinant DNA molecule, the "*replicon*;"— the cell will not replicate any foreign DNA in the absence of a suitable vector.

  In our application, BACs will be used more or less as a measuring device. If two probes cohybridize to the same BAC then we know that those two probes are within a distance smaller than the length of the BAC. But, just hybridizing with one BAC at a time will be inefficient; with further analysis, we will see that hybridizing with a several thousands randomly selected BACs can give us distance information for many pairs of probes, simultaneously. The fact that we can make vast amount of copies of the same BAC reliably and rapidly, is the key to the overall robustness of our approach.

  PCR or polymerase chain reaction is an *in vitro* technique used to replicate a fragment of DNA so as to produce many copies of a short specific DNA sequence. The biochemical process involved in PCR operates iteratively: in one step, two strands of the DNA are denatured (separated) by heating, and in the subsequent step, short sequences of a single DNA strand (primers) are added,

together with a supply of free nucleotides and DNA polymerase, to create two double stranded copies each originating from the two complementary single strands obtained in the earlier step. The original DNA sequence doubles in each repetition of the heating and cooling cycle and results in rapid amplification. PCR is commonly used as an alternative to *in vivo* cloning as a means for amplifying DNA material. PCR is used in many medical and biological applications (measuring gene expressions, DNA sequencing etc.), but has found its most prominent applications in forensic science as a tool for amplifying minuscule traces of genetic material needed for DNA fingerprinting.

4 Probes and their Distances

We will rely on the available microarray technology to assist us in measuring the pair-wise distances among a large number of probes. As will be explained shortly, the basic technology uses unordered probes that are microarrayed and hybridized to an organized sampling of arrayed but unordered members of libraries of large insert genomic clones (e.g. BAC, Bacterial Artificial Chromosomes). The basic ideas of this process can be further generalized with other types of clones, chromosomal fragments or random PCR products derived from genomic DNA.

In order to completely appreciate the challenges and the full potential of this technology, a detailed discussion must include our knowledge of genome organization, DNA hybridization, repetitive DNA, gene duplication, and the varieties of microarrays. But, for the sake of simplicity, we omit these details.

Imagine a set of P points on a line segment of length G (e.g. probes on a chromosome or a genome, which denotes the collection of all the chromosomes) and a set of random intervals of length L from the line segment (e.g. a BAC or YAC library, or the chromosomal fragments contained in a panel of radiation hybrid cell lines). For our purposes, these line segments will be BACs and the length $L = 160Kb$. We perform the following "array hybridization." We pick two random subsets of K intervals each and denote one set as the *red* set and the other as the *green* set. We assign each point a color: "B=blank" ($\neg$Red $\wedge$ $\neg$Green), "R=red (Red $\wedge$ $\neg$Green), "G=green ($\neg$Red$\wedge$Green), or "Y=yellow" (Red$\wedge$Green), based on whether the point belongs to neither the union of intervals in the red set nor the union of intervals in the green set (blank), the former (red), the later (green) or the both(yellow). These logical steps are easily achieved by an "array hybridization" step with microarray. The P probes are Watson-Crick complements of short "unique" subsequences of the genomes and can be produced reliably and in large quantity with the use of restriction enzymes, or be synthesized as oligoes. Each probe is spotted at a fixed physical location on a microarray. Now, if a collection of several BACs are hybridized to this microarray, those BACs that contain a subsequence, complementary to the probe-sequence, hybridize to the probe. Since these BACs possess a color (physically achieved by attaching a colored fluorescent dye), the probe acquires the colors of the BACs that it hybridizes to. For instance, if the complement of the probe-sequence is contained in a BAC sequence, dyed red, but not in any BAC sequence, dyed green, then that probe will be seen red. Analogously, the relation between points

and intervals in our discussion earlier, can be seen to be biochemically determined for the probes and BACs through hybridization. Thus array hybridization allows us to observe a color outcome for each of the 150,000 probes in a short constant amount of time.

Notice that the probability that two probes have different color outcomes in a single array hybridization, depends on how far apart they are and monotonically increases with the distance. Thus, if we can estimate this probability by several (M) array hybridization experiments then we can estimate the distance between two probes. The probability is easily estimated by counting the number of experiments in which the probes have different color outcomes and expressing it as a fraction of the total number of experiments. In other words, we can present the outcomes of M different experiments as "color-vectors" of length M, one associated to each probe, and estimate the distance between two probes from the Hamming distance between their associated color vectors. Note that the Hamming distance between two discrete-valued vectors is defined as the the number of positions where the entries of the two vectors differ.

In order to explore the relation between the "true" distance between probes and the Hamming distance between their color vectors, we proceed as follows: Represent the probes as points $\{p_1, \ldots, p_P\}$. Assume that the probes are i.i.d. with uniform random distribution over the interval $[0, G]$. Let S be a collection of intervals of the genome, each of length L. Suppose the left-hand points of the intervals of S are i.i.d. uniform random variables over the interval $[0, G]$. Take a small subset (of size $2K$) of intervals $S' \subset S$, chosen randomly from S. Divide S' randomly into two equal-sized disjoint subsets $S' = S'_R \cup S'_G$, where R indicates a red color set and G indicates a green color set. Now specify any point p_i in $[0, G]$ and consider the possible associations between p_i, and the intervals in S':

- Point p_i is not covered by any interval in S'. Probe p_i hybridizes to zero BACs. We say the outcome is 'B' (blank).
- Point p_i is covered by at least one interval of S'_R but no intervals of S'_G. Probe p_i hybridizes to at least one red BAC and zero green BACs. We say the outcome is 'R' (red).
- Point p_i is covered by at least one interval of S'_G but no intervals of S'_R. Probe p_i hybridizes to at least one green BAC and zero red BACs. We say the outcome is 'G' (green).
- Point p_i is covered by at least one interval of S'_R and at least one interval of S'_G. Probe p_i hybridizes to at least one green BAC and at least one red BAC. We say the outcome is 'Y' (yellow).

We call these events i_B, i_R, i_G, and i_Y respectively. If we perform a sequence of M such experiments then for each p_i we get a sequence of M outcomes represented as a color vector of length M. The parameter domain for the full experiment is $\langle P, L, K, M \rangle$, where P is the number of probes, L is the average length of the genomic material used (for BACs, $L = 160$kb), K is the sampling size, and M is the

number of samples. The output is a color sequence for each probe. The sequence corresponding to probe p_j is $\mathbf{s}_j = \langle s_{j,k} \rangle_{k=1}^{M}$ with $s_{j,k} \in \{B, R, G, Y\}$.

With the resulting color sequences $\mathbf{s}_j$ we can compute the pairwise Hamming distance. Let

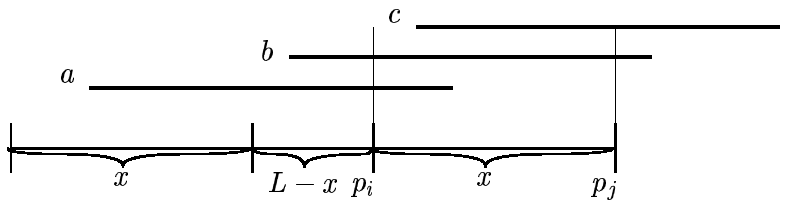
$$\begin{aligned} H_{i,j} &= \#\text{ places where } \mathbf{s}_i \text{ and } \mathbf{s}_j \text{ differ,} \\ C_{i,j} &= \#\text{ places where } \mathbf{s}_i \text{ and } \mathbf{s}_j \text{ are the same but } \mathbf{s}_i \neq B, \\ T_{i,j} &= \#\text{ places where } \mathbf{s}_i \text{ and } \mathbf{s}_j \text{ are } B. \end{aligned}$$

Note that the Hamming distance $H_{i,j}$ defines a distance metric on the set of probes. The roles of the functions $C_{i,j}$ and $T_{i,j}$ will become clear, as we go on.

Since the M array hybridization experiments are independent, we need to look at any single experiment, i.e., $M = 1$ case. Let us define events $T = (i_B \wedge j_B)$, $C = ((i_R \wedge j_R) \vee (i_G \wedge j_G) \vee (i_Y \wedge j_Y))$, and $H = (\neg T \wedge \neg C)$. We will compute the conditional probabilities of these events when we know the distance between the corresponding probes, i.e., $x = |p_i - p_j|$.

Given a set of $2K$ BACs on a genome $[0, G]$ the probability that none start in an interval of length l is $(1-\alpha)^l \approx e^{-\alpha l}$ where $\alpha = \frac{2K}{G}$. Similarly, the probability that no red (respectively, green) BACs start in an interval of length l is $(1-\alpha_R)^l \approx e^{-\alpha_R l}$ (respectively, $e^{-\alpha_G l}$) where $\alpha_R = \alpha_G = \frac{K}{G} = \alpha/2$. Let c denote $\alpha L = 2KL/G$, the coverage by the BAC sublibrary $S' \subset S$.

Shown below is a diagram that is helpful in computing the probabilities for events C, H and T when $x < L$. The heavy dark bar labeled a represents a set of BACs which covers probe p_i but not p_j; the bar labeled b represents a set of BACs that covers probe p_i and p_j; finally, the bar labeled c represents a set of BACs that covers p_j but not p_i.



Hence we can compute various conditional probabilities:

$$\begin{aligned} P(T|x \leq L) &= e^{-(\alpha_R+\alpha_G)(L+x)} \\ P(i_R \wedge j_R|x < L) &= e^{-\alpha_G(L+x)}\{1 - 2e^{-\alpha_R L} + e^{-\alpha_R(L+x)}\} \\ P(i_G \wedge j_G|x \leq L) &= e^{-\alpha_R(L+x)}\{1 - 2e^{-\alpha_G L} + e^{-\alpha_G(L+x)}\} \\ P(i_Y \wedge j_Y|x \leq L) &= (1 - 2e^{-\alpha_R L} + e^{-\alpha_R(L+x)}) \\ &\quad \times (1 - 2e^{-\alpha_G L} + e^{-\alpha_G(L+x)}) \\ P(C|x \leq L) &= P(i_R \wedge j_R|x \leq L) \\ &\quad + P(i_G \wedge j_G|x \leq L) \end{aligned}$$

$$
\begin{aligned}
& + P(i_Y \wedge j_Y | x \leq L) \\
P(H|x \leq L) &= 1 - [P(T|x \leq L) + P(C|x \leq L)]
\end{aligned}
$$

Similarly, when $x \geq L$ the probabilities are:

$$
\begin{aligned}
P(T|x \geq L) &= e^{-(\alpha_R + \alpha_G)(2L)} \\
P(i_R \wedge j_R | x \geq L) &= e^{-\alpha_G(2L)}\{(1 - e^{-\alpha_R L})^2\} \\
P(i_G \wedge j_G | x \geq L) &= e^{-\alpha_R(2L)}\{(1 - e^{-\alpha_G L})^2\} \\
P(i_Y \wedge j_Y | x \geq L) &= (1 - e^{-\alpha_R L})^2 (1 - e^{-\alpha_G L})^2 \\
P(C|x \geq L) &= P(i_R \wedge j_R | x \geq L) \\
&\quad + P(i_G \wedge j_G | x \geq L) \\
&\quad + P(i_Y \wedge j_Y | x \geq L) \\
P(H|x \geq L) &= 1 - [P(T|x \geq L) + P(C|x \geq L)]
\end{aligned}
$$

Recall that $\alpha_R L = \alpha_G L = \frac{c}{2} = \frac{KL}{G}$. Let $q = q(x) = P(H)$ and $p = p(x) = P(C)$. In general $q(x)$ and $p(x)$ are complicated functions of x, shown below:

$$
\begin{aligned}
q(x) = P(H) &= \frac{2c\exp(\frac{-c}{2})x}{L} + O(x^2) \\
p(x) = P(C) &= 1 - e^{-c} + \frac{c}{2}(e^{-c} - 2e^{-\frac{c}{2}})x + O(x^2)
\end{aligned}
$$

With independent sampling, we now have the following Binomial probability distribution functions:

$$
\begin{aligned}
P(H_{i,j}) &\sim \text{Binomial}(M,\ q(x)) \\
P(C_{i,j}) &\sim \text{Binomial}(M,\ p(x))
\end{aligned}
$$

Solving the equations above, we have

$$
\tilde{x} \approx \left(\frac{q}{q + 2p} e^{c/2}\right) L.
$$

We can use the following estimator of x_{ij} to measure the distance between two probes:

$$
\tilde{x}_{ij} = \frac{H_{i,j}}{H_{i,j} + 2C_{i,j}} e^{\frac{H_{i,j} + 2C_{i,j}}{4M}} L.
$$

Note that this estimator takes into account the variation of sample coverage over the genome. Using a simplifying normal approximation, we have, for $x < L$, the measured distance $\tilde{x}$:

$$
\tilde{x} \sim x + \left(\frac{e^{c/4}}{\sqrt{2c}}\right) \sqrt{\frac{L}{M}} \sqrt{x} \mathbf{N}(0, 1).
$$

When $x \geq L$, similarly we have:

$$\tilde{x} \sim L + \left(\frac{e^{c/4}}{\sqrt{2c}}\right) L\sqrt{\frac{1}{M}}\mathbf{N}(0,1).$$

Here $\mathbf{N}(0,1)$ represents a standard normal distribution of mean 0 and variance 1.

In summary, our biochemical process provides a way of measuring the distance between any two probes. Furthermore, we have a good model of the errors in the measurements and we can accurately control the amount of error by appropriately choosing various experimental parameters such as K = the number of BACs (this affects the parameter c), L = the clone length and M = number of array hybridization experiments. Also, we should note that if two probes are further than the BAC length (i.e., $L = 160Kb$), the distance measured does not provide any useful information

5 Applications

Finally, we need to see how to use the probe-distance technology to compare two genomes.

In the simplest possible applications, the probe-distance data can be used to find the relative locations of the probes along the genome. The information created this way provides us a low-resolution reference map of the probes. Now, a specific genome (e.g., from tumor tissues) can be compared with this map to see which of these probes are present multiple times and which probes are deleted. The simplest analysis could involve hybridization with whole genomic DNA to microarrays of probes. If a region surrounding a probe is missing from the selected genome, then the genomic DNA lacks material that could hybridize to the probe. Conversely, if a certain region surrounding a probe has been amplified in the selected genome, then the genomic DNA has material that could hybridize to the probe in abundance. Thus, such an analysis employed with cancer genomes can tell us the regions of amplification and deletion, but not translocations. This analysis, nonetheless, would be sufficient to find the oncogenes and tumor suppressor genes.

But, while the ideas described in the preceding paragraph are, in principle, sound, they are impractical, since the complexity of the genome is high and the signal to noise ratio is inadequate to detect all but the grossest amplifications. There have been many interesting modifications to the basic technology, in the form of "representations resulting in complexity reduction," that have improved the signal to noise ratio and detected the copy-number changes accurately. (Amplifications and deletions are specific examples). (See [4, 6, 7, 8])

Further improvements to the basic technology are achieved when the probe-distances are measured with genomic chromosomal fragments, instead of the clones. When clones from a library are used, the distances measured are distances with respect to a reference genome and depends on how the clone library was created. If the clones are avoided and genomic materials from a selected genome are used to measure the distances between probe pairs, then the measured distances reflect the

locations of the probes along the selected genome and hence, much more informative. As before, the signal to noise ratio in the hybridization creates problems and can be solved by various modifications to the basic technology.

In general, comparative genomics has many applications of utmost biological significance and the technology developed here can be adapted to many different applications in those contexts. But most importantly, the ideas developed here indicate how the design principles, developed for computer algorithms, information theory and systems sciences, etc., are likely to find applications in biotechnology. The biggest impact of biocomputing will be in biotechnology.

6 Bibliographic Notes

The basic ideas of the algorithm described here and their extension to create genome-wide maps of probes can be found in the paper by Casey, Mishra and Wigler [4]. The experimental work as well as the underlying foundations for detecting gene copy number fluctuations are to be found in Lucito et al. [7]. The other related ideas (e.g., low complexity representation of genomes, cloning genomic differences, application to genetic analysis, etc.) are described in [6, 8, 7]. The algorithms and algorithmic complexity of constructing probe maps, RH maps and similar physical maps are discussed in [1, 2, 5, 9, 10]. A good reference for the biotechnology revolution spurred by the human genome project is the recent book of Cantor and Smith [3].

References

[1] F. Alizadeh, R.M. Karp, D.K. Weisser, and G. Zweig. "Physical Mapping of Chromosomes Using Unique Probes," **Journal of Computational Biology, 2(2)**:159–185, 1995.

[2] A. Ben–Dor, and B. Chor. "On constructing radiation hybrid maps," **Proceedings of the First International Conference on Computational Molecular Biology,** 17–26, 1997

[3] C. Cantor, and C. Smith. **Genomics: The Science and Technology Behind the Human Genome Project**,
John Wiley and Sons, New York, 1999.

[4] W. Casey, B. Mishra, and M. Wigler. "Placing Probes along the Genome using Pair-wise Distance Data," In *Algorithms in Bioinformatics, First International Workshop, WABI 2001 Proceedings*, **LNCS 2149**:52-68, Springer-Verlag, 2001.

[5] J. Håstad,L. Ivansson,J. Lagergren "Fitting Points on the Real Line and its Application to RH Mapping," **Lecture Notes in Computer Science, 1461**:465–467, 1998.

[6] N. Lisitsyn,and M. Wigler "Cloning the differences between two complex genomes," **Science, 258**:946–951,1993.

[7] R. Lucito, J. West, A. Reiner, J. Alexander, D. Esposito, B. Mishra, S. Powers, L. Norton, and M. Wigler "Detecting Gene Copy Number Fluctuations in Tumor Cells by Microarray Analysis of Genomic Representations," **Genome Research, 10(11)**: 1726–1736, 2000.

[8] R. Lucito, M. Nakimura, J. West, Y. Han, K. Chin, K. Jenson, R. McCombie, and M. Wigler "Genetic analysis using genomic representations," **Proc. Natl. Acad. Sci. USA, 95**:4487–4492, 1998.

[9] M. Jain, and E.W. Myers. "Algorithms for Computing and Integrating Physical Maps Using Unique Probes," **Journal of Computational Biology, 4(4)**:449–466, 1997.

[10] D. Slonim, L. Kruglyak, L. Stein, and E. Lander "Building human genome maps with radiation hybrids," **Journal of Computational Biology, 4(4)**:487–504, 1997.