Journal of Bioinformatics and Computational Biology
Vol. 4, No. 3 (2006) 639–647
© Imperial College Press

# A NOTE ON PHASING LONG GENOMIC REGIONS USING LOCAL HAPLOTYPE PREDICTIONS

ELEAZAR ESKIN
*Computer Science and Engineering, University of California, San Diego*
*eeskin@cs.ucsd.edu*

RODED SHARAN
*School of Computer Science, Tel-Aviv University*
*Tel-Aviv 69978, Israel*
*roded@post.tau.ac.il*

ERAN HALPERIN
*International Computer Science Institute, Berkeley*
*heran@icsi.berkeley.edu*

Received 21 August 2005
Revised 22 October 2005
Accepted 13 November 2005

The common approaches for haplotype inference from genotype data are targeted toward phasing short genomic regions. Longer regions are often tackled in a heuristic manner, due to the high computational cost. Here, we describe a novel approach for phasing genotypes over long regions, which is based on combining information from local predictions on short, overlapping regions. The phasing is done in a way, which maximizes a natural maximum likelihood criterion. Among other things, this criterion takes into account the physical length between neighboring single nucleotide polymorphisms. The approach is very efficient and is applied to several large scale datasets and is shown to be successful in two recent benchmarking studies (Zaitlen *et al.*, in press; Marchini *et al.*, in preparation). Our method is publicly available *via* a webserver at http://research.calit2.net/hap/.

*Keywords*: Haplotype phasing; SNPs.

## 1. Introduction

Single nucleotide polymorphisms (SNPs) are differences, across the population, in a single base, within an otherwise conserved genomic sequence. Approximately, 10 million common SNPs,[3,4] each with a frequency of 10–50%, account for the majority of the variation between DNA sequences of different people.[5] The variation in the allelic content of SNPs may be associated with medical conditions. Thus, efficient and accurate methods for obtaining SNP information are of great clinical, scientific, and commercial value.

The sequence of alleles in contiguous SNP positions along a chromosomal region is called a *haplotype*. For diploid organisms, two haplotypes make up a *genotype*, which is the list of allele-pairs along the chromosomal segment. The genotype contains information solely on the combination of alleles in a given site and not on the association of each allele with one of the two chromosomes, also called its *phase*. The current technology, suitable for large-scale polymorphism screening, obtains the genotype information at each SNP, but not its phase. The latter information can be obtained at a considerably higher cost.[5] It is, therefore, desirable to develop efficient methods for inferring haplotypes from genotype information.

Numerous approaches have been suggested in the literature to resolve haplotypes from genotype data. These methods include the seminal approach of Clark,[6] parsimony approaches,[7–9] maximum likelihood methods,[10–13] statistical methods such as PHASE[14] and HAPLOTYPER,[15] and perfect phylogeny-based approaches.[16,17] All these methods perform very well across short genomic regions with limited diversity (see Fig. 1), but a few extend to large regions with high diversity. Consider, for example, the entire 103 SNPs in the 616 kB region examined in Ref. 18. Out of the 258 haplotypes in the population, the most common haplotype only occurs in 45 individuals and 169 haplotypes occur only in one individual. As an effort to characterize human variation will be a tremendous undertaking,[3] methods for haplotyping long genomic regions will be essential for analyzing data from large-scale genotype studies, including data generated for whole genome association studies.[19,20]

In this paper, we describe a novel method, HAP-TILE, for combining local phasing predictions for phasing long genomic regions. Our method is based on using accurate phase predictions over short overlapping regions, obtained by any extant method, to recover haplotypes over long regions. We present an efficient dynamic programming algorithm for optimally combining the overlapping local predictions with respect to a natural maximum likelihood criterion. The maximum likelihood criterion takes into account an estimate of the accuracy of the prediction based on the physical length of the region and the entropy of the distribution of the haplotypes therein.

Our method follows similar intuitions to the partition-ligation (PL) method, which was used in HAPLOTYPER[15] and subsequently in PL-EM.[21] In PL method, a long region is partitioned into a set of short regions' each of the regions is phased, and neighboring regions are then phased together recursively until a complete haplotype is reconstructed. One deficiency with the PL method is that the short regions are chosen arbitrarily, and due to the nature of the ligation step, the method is not guaranteed to produce a global optimum. In contrast, our method considers predictions over all possible short region segments, and uses a tiling technique which is guaranteed to find a solution with maximum likelihood. The latest version of PHASE[22] uses a different model for combining local predictions into a longer haplotype prediction.

The main advantage of HAP-TILE is the efficiency of the algorithm for combining local predictions into longer regions. When coupled with the HAP algorithm for making local predictions,[23] the approach is a factor of a 1000 times faster than PHASE as verified in two recent benchmarking studies.[1,2] The speed of the algorithm is both due to the speed of HAP and the efficiency of the dynamic programming algorithm. The speed of the algorithm allowed it to be applied to whole genome datasets such as the Perlegen data set which contains over 1.5 million SNPs in 71 individuals for a total of over 100 million genotypes[19] as well as the entire contents of the genotype portion of the NCBI dbSNP database, which contains over 286 million genotypes.[1] The benchmarking studies have shown that HAP-TILE using HAP for local predictions has comparable accuracy to other state of the art methods.[1,2] HAP-TILE, using HAP for making local predictions, is publicly available *via* a webserver at http://research.calit2.net/hap/.

The rest of the paper is organized as follows: Section 2 presents our probabilistic model for local haplotype predictions over a given region, and the computational problem of computing a maximum likelihood solution to the haplotyping problem under this model. Section 3 studies the complexity of the latter problem and gives a dynamic programming solution for it. Finally, Section 4 details the steps of our practical haplotyping algorithm.

## 2. The Generative Probabilistic Model

In this section, we define a probabilistic model for the generation of local predictions of phasing algorithms given a set of genotypes over some genomic region. We focus on binary SNPs (having only two alleles). We use the following notation: a haplotype $H$ is a binary string. A genotype $G$ is a string over the alphabet $\{0, 1, 2\}$. We say that a genotype $G \in \{0, 1, 2\}^n$ is *compatible* with the haplotypes $H_1, H_2 \in \{0, 1\}^n$, if for every $i$ the following two conditions hold: (1) if $G(i) = 1$ or $G(i) = 0$, i.e. $i$ is a *homozygous* position, then $H_1(i) = H_2(i) = G(i)$; and (2) if $G(i) = 2$, i.e. $i$ is an *heterozygous* position, then $H_1(i) \neq H_2(i)$. If $H_1, H_2$ are compatible with $G$, we say that $(H_1, H_2)$ is a *phase* of $G$.

Let $G_1, \ldots, G_t$ be the input genotypes, where the (true) phase of $G_i$ is $(F_i^*, M_i^*)$. We consider $(n-k+1)$ windows, $W_0, W_1, \ldots, W_{n-k}$, each of length $k$, where window $W_l$ contains positions $l+1, \ldots, l+k$. For every genotype $G_i$, and every window $W_l$, the model generates two haplotypes $H_{0l}^i, H_{1l}^i \in \{0, 1\}^k$ consistent with $G_i$ in window $W_l$, which we call the *local predictions* of window $W_l$. At first, $H_{0l}^i(j) = M_i^*(l+j)$ and $H_{1l}^i(j) = F_i^*(l+j)$, that is, $H_{0l}^i$ and $H_{1l}^i$ are simply the copies of the two haplotypes in those positions. We then swap the values of $H_{0l}^i$ and $H_{1l}^i$ with probability $\frac{1}{2}$. Therefore, the resulting haplotypes satisfy that with probability $\frac{1}{2}$, $H_{0l}^i$ is a copy of $F_i^*$ (in the corresponding positions) and $H_{1l}^i$ is a copy of $M_i^*$, and with probability $\frac{1}{2}$ it is the other way around. Finally, we independently swap the values of $H_{0l}^i(j)$ and $H_{1l}^i(j)$ with probability $p < \frac{1}{2}$ for every position $1 \leq j \leq k$.

Suppose now that $H^i_{0l}, H^i_{1l}$ are local predictions for the genotypes, generated as described above, where $i = 1, \ldots, t$ and $l = 0, \ldots, n-k$. Let $(F_1, M_1), \ldots, (F_t, M_t)$ be a suggested phasing of the genotypes. Then, the log likelihood of this solution according to our model is:

$$L = \sum_{i=1}^{t} \sum_{l=0}^{n-k} \left[ \min\{h^i_{0l}, h^i_{1l}\} \log \frac{p}{1-p} + k \log(1-p) \right],$$

where $h^i_{bl}$, for $b = 0, 1$, is the total number of disagreements between $H^i_{bl}$ and $F^i$ and between $H^i_{(1-b)l}$ and $M^i$, at positions $l+1, \ldots, l+k$.

Our goal is to find a solution with maximum likelihood. Since the likelihood function decomposes over the individuals, we can maximize it separately for each individual. For the $i$th individual, this amounts to finding a pair of haplotypes $(F^i, M^i)$, for which $\sum_{l=0}^{n-k} \min\{h^i_{0l}, h^i_{1l}\}$ is minimized. This gives rise to the following problem:

*Problem* 1 (*Minimum Conflict Phasing* (*MCP*)). Given an unphased genotype $G$ and a set of local prediction for it, each of which is compatible with $G$, find two haplotypes that are compatible with $G$ and minimize the number of disagreements with the local predictions.

## 3. The Minimum Conflict Phasing Problem

In this section, we study the Minimum Conflict Phasing problem. First, we prove that the problem is NP-hard. We then provide a linear time algorithm for it, when the length of a local prediction is fixed.

**Theorem 1.** *Minimum Conflict Phasing is NP-hard.*

**Proof.** We give a reduction from MAX-CUT. Let $\langle K = (V, E), r\rangle$ be an instance of MAX-CUT. Define an instance of MCP as follows: we set the window length $k$ to $|V| + 2|E|$, and the length of the genotype $n$ to $|V| + 4|E| - 1$. Thus, the total number of windows is $n - k + 1 = 2|E|$. We let $P = \{2|E|+1, \ldots, 2|E| + |V|\}$ be the set of positions shared by all windows, which we call *vertex positions*. For convenience, we refer to position $2|E| + i$, corresponding to vertex $i \in V$, as $v_i$. We define the genotype $G$ as having missing entries over all vertex positions, and being homozygous with a value of 1 elsewhere. With every edge $e \in E$, we associate two arbitrary windows $W_e, W'_e$. If $e = (i, j)$, the local predictions for the two windows $W_e, W'_e$ are set in the following way: let $H_1, H_2$ and $H'_1, H'_2$ be the two pairs of haplotypes corresponding to the two windows. For positions $v_i, v_j$ we set $H'_1(v_i) = H_1(v_i) = 0$, $H_1(v_j) = H'_1(v_j) = 1$, $H'_2(v_i) = H_2(v_i) = 1$ and $H_2(v_j) = H'_2(v_j) = 0$. For every other vertex position $l$, we set $H_1(l) = H_2(l) = 0$ and $H'_1(l) = H'_2(l) = 1$. In every nonvertex position, all windows are homozygous with value 1. It should be noted that the resulting local predictions are consistent with the genotype $G$, since every vertex position is missing in $G$, and every other position is homozygous 1 both in $G$ and in the local predictions.

We now claim that $K$ has a maximum cut with size at least $r$, if and only if the MCP instance has a solution with at most $2(|E||V| - 2r)$ disagreements. Suppose there is a phase of $G$ with at most $2(|E||V| - 2r)$ disagreements. In particular, consider a phase $(F, M)$ that induces a minimum number of disagreements. We first claim that without loss of generality, for every vertex position $v_i$ we have $F(v_i) \neq M(v_i)$. Consider the optimal solution with the minimum number of homozygous positions (that is, positions where $F(v_i) = M(v_i)$). We will show that this solution actually has no homozygous vertices. Otherwise, there is a homozygous vertex $v_i$ such that $F(v_i) = M(v_i)$. Consider the following two alternative solutions $(F_1, M_1)$ and $(F_2, M_2)$. These solutions are identical to $(F, M)$ on every position, except $i$. In position $i$, $F_1(v_i) = M_2(v_i) = 1$ and $M_1(v_i) = F_2(v_i) = 0$. For every edge $(i, j)$, the number of conflicts of $(F, M)$ at position $i$ is exactly two (i.e. out of $H_1, H_2, H_1', H_2'$ exactly two have the value $F(v_i)$ at position $i$. On the other hand, the average number of conflicts of $(F_1, M_1)$ and $(F_2, M_2)$ is also two. Thus, summing over all edges $(i, j)$, the average number of conflicts of $(F_1, M_1)$ and $(F_2, M_2)$ is the same as the number of conflicts of $(F, M)$. Therefore, one of the alternative solutions is optimal, since the number of conflicts for that solution is at most the number of conflicts of $(F, M)$. At the same time, the alternative solution has one less homozygous vertex, contradicting the definition of $(F, M)$ as an optimal solution with a minimal number of homozygous positions.

Consider the cut induced by the set $S = \{i \in V \mid F(v_i) = 1\}$ of vertices, whose corresponding positions were assigned 1 in $F$. Let $s$ denote the number of edges crossing the cut. For every edge $(l_1, l_2) \in E$, if $F(v_{l_1}) \neq F(v_{l_2})$ then the number of conflicts with the windows $W_e, W_e'$ in positions $v_{l_1}$ and $v_{l_2}$ is zero. If $F(v_{l_1}) = F(v_{l_2})$, then the number of disagreements is four. For every other vertex position, the number of conflicts with $W_e, W_e'$ is exactly two, and for every nonvertex position, the number of conflicts with $W_e, W_e'$ is zero. Therefore, the total number of conflicts is

$$4|\{(l_1, l_2) \in E | F(v_{l_1}) = F(v_{l_2})\}| + 2(|V| - 2)|E| = 2|E||V| - 4s \leq 2|E||V| - 4r.$$

Conversely, given a cut $(S, \bar{S})$ of size at least $r$, we define $F$ to have value 1 in nonvertex positions. For a vertex position $v_i$ we define $F(v_i) = 1$, if and only if $i \in S$, and $M(v_i) = 1$, if and only if $i \notin S$. It is easy to verify that the number of disagreements induced by this solution is at most $2|E||V| - 4r$. □

### 3.1. *A dynamic programming solution*

We now provide a linear time dynamic programming solution to MCP, when the size of the window $k$ is fixed. We assume that we are given a genotype $G$ of length $n$, and local predictions $H_{0l}, H_{1l}$ for $0 \leq l \leq n - k$. In what follows, we describe the construction of one of the haplotypes $F$. The other haplotype $M$ can be derived from $F$ and $G$ in a straightforward manner.

Denote by $S(j, r)$ the best haplotype assignment for the first $j + k$ positions in $F$, where the last $k$ bits are $r = r_1, \dots, r_k$. For every assignment $r = r_1, \dots, r_k$ to $F$ at positions $j + 1, \dots, j + k$, denote by $h_b(j, r)$, the total number of disagreements between $H_{bj}$ and $r$ and between $H_{(1-b)j}$ and $\bar{r}$, where $\bar{r}$ is the implied assignment to $M$ at these positions. Let $h(j, r) = \min\{h_0(j, r), h_1(j, r)\}$. Then, the following recurrence formula gives $S(j + 1, r)$:

$$S(j+1, (r_1, \dots, r_k)) = \min_{b=0,1} \{S(j, (b, r_1, \dots, r_{k-1})) + h(j+1, (r_1 \dots r_k))\},$$

where $S(0, r) = h(0, r)$ for all $r$. It is easy to compute $h(j, (r_1 \dots r_k))$ for every possible $j$ and $r$ in time $O(2^k n)$. Using the recurrence formula, we can find $S(n-k, r)$ for all $r$. By tracing the solution which leads to a minimal value of $S(n - k, r)$ (over all values of $r$), we can reconstruct the haplotypes that attain the maximum likelihood.

## 4. The Practical Algorithm

We devised a three-step method, called HAP-TILE, for phasing genotype data, which is based on the dynamic programming algorithm presented in Sec. 3.1. HAP-TILE starts by computing local predictions for all possible short segments of the genotyped region (up to length 12). Then, confidence scores are assigned to each local prediction. Finally, the dynamic programming algorithm is used to tile the local predictions into complete haplotype predictions.

We scan the genotypes with a sliding window and compute the local predictions in each window. In practice, we do not use a fixed-size window, but rather use all possible window sizes from 2 to $L$ (where $L = 12$). This is needed, since the density of heterozygous SNPs may vary considerably along the typed region. Hence, at every SNP $j$, we have $L - 1$ local predictions starting at this SNP.

To make the local predictions, we apply the HAP algorithm to each local region.[23] The HAP algorithm builds on perfect phylogeny principles, and assumes that the set of haplotypes satisfies the four gamete test, that is, at most three allele combinations are observed for any pair of marker positions.[16]

With each local prediction, we associate a confidence level $p(j, k)$, which reflects the probability that a local prediction of length $k$ that starts at SNP $j$ is correct. The estimation of these confidence levels assumes that the less diverse the haplotypes in a region are, the more accurate their prediction will be (see Fig. 1). We compute a confidence level as the product of two figures. The first is an *a priori* estimate of the probability of having strong correlation in a certain region based on its physical length. Let $l(j, k)$ be the distance between SNPs $j$ and $j+k$. The prior is based on the length of the region $p_l(j, k) = \exp\left(-\frac{l(j,k)}{R}\right)$, where $R$ is a parameter which controls the decay of the confidence in the prediction with length. We use an exponential distribution for this estimate, as commonly used for modeling the occurrence of recombinations. This allows us to take into account the distance between SNPs in our predictions. The second figure $p_e(j, k)$ is an estimate of the probability of

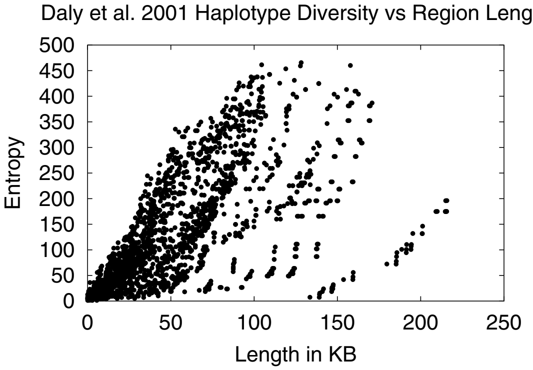


Fig. 1. Haplotype diversity as a function of region length for the data of Daly *et al.*[18] Each point corresponds to a region. The $x$-axis shows the length of the region in kilobases and the $y$-axis shows the entropy of the haplotype distribution. For shorter regions, the entropy of the distribution is smaller, and the haplotypes are less diverse demonstrating that shorter haplotype blocks are likely to be more accurate.

having such a phase prediction given that the data is generated by random mating of individuals from the population, whose sample is observed by $u_0$. This estimate is computed as in Ref. 17. This in turn, can be shown to be equivalent to the entropy of the haplotype distribution.

In order to combine the estimated confidence levels into the dynamic programming algorithm, we redefine $h(j,r)$ as follows: using the notation of Sec. 3.1, let $h_b^i(j,r)$ be the total number of disagreements for a prediction of length $i$. We define:

$$h(j,r) \equiv \sum_{i=2}^{L} p(j,i) \min\{h_0^i(j,r), h_1^i(j,r)\},$$

where $p(j,k) = p_l(j,k)p_e(j,k)$.

## 5. Conclusions

Recent studies on haplotype structure have shown that haplotypes have limited diversity in local regions. In these regions, many methods can resolve the haplotypes from genotype data collected for a population. We have presented HAP-TILE, a method for combining local haplotype predictions into longer haplotypes. HAP-TILE coupled with local haplotype prediction algorithm HAP has been applied to the Perlegen whole genome variation dataset[19] and the entire genotype portion of the NCBI database.[1] The efficiency of the approach made possible the inference of haplotypes in these large datasets. Two recent benchmarking studies have shown that HAP-TILE using HAP for local predictions has comparable accuracy to other state of the art methods.[1,2]

## Acknowledgments

We thank Richard Karp for helpful discussions. E. E. is supported by the California Institute for Telecommunications and Information Technology (Calit2). This research was supported by NSF grant SEIII-0513612. R. S. was supported by an Alon fellowship.

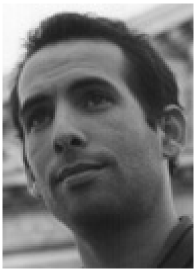
## References

1. Zaitlen N, *et al.*, Inference and analysis of haplotypes from combined genotyping studies deposited in dbSNP, *Genome Res* (in press.), 2005.
2. Marchini J, *et al.*, A comparison of phasing algorithms for trios and unrelated individuals, in preparation, 2005.
3. NIH: Large-scale genotyping for the haplotype map of the human genome, RFA: HG-02-005, 2002.
4. Group TISMW, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature* **409**:928–33, 2001.
5. Patil N, Berno A, Hinds D, *et al.*, Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, *Science* **294**:1719–23, 2001.
6. Clark A, Inference of haplotypes from PCR-amplified samples of diploid populations, *J Mol Biol Evol* **7**:111–122, 1990.
7. Gusfield D, A practical algorithm for optimal inference of haplotypes from diploid populations, *Proc Eighth Int Conf Intell Syst Mol Biol*, pp. 183–189, 2000.
8. Gusfield D, Inference of haplotypes from samples of diploid populations: complexity and algorithms, *J Comput Biol* **8**:305–23, 2001.
9. Lancia G, *et al.*, SNPs problems, algorithms and complexity, european symposium on algorithms, in Springer (ed.), *Proc Eur Symp Algorithms (ESA'01)*, Lecture Notes in Computer Science, Vol. 2161, pp. 182–193, 2001.
10. Excoffier L, Slatkin M, Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, *Mol Biol Evol* **12**:921–7, 1995.
11. Fallin D, Schork N, Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data, *Am J Hum Genet* **67**:947–959, 2000.
12. Hawley M, Kidd K, Haplo: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes, *J Heredity* **86**:409–11, 1995.
13. Long J, Williams R, Urbanek M, An EM algorithm and testing strategy for multiple-locus haplotypes, *Am J Hum Genet* **56**:799–810, 1995.
14. Stephens M, Smith N, Donnelly P, A new statistical method for haplotype reconstruction from population data, *Am J Hum Genet* **68**:978–989, 2000.
15. Niu T, *et al.*, Bayesian haplotype inference for multiple linked single nucleotide polymorphisms, *Am J Hum Genet* **70**:157–169, 2002.
16. Gusfield D, Haplotyping as perfect phylogeny: conceptual framework and efficient solutions, *Proc 6th Int Conf Comput Mol Biol (RECOMB'02)*, pp. 166–175, 2002.
17. Eskin E, Halperin E, Karp R, Efficient reconstruction of haplotype structure via perfect phylogeny, *J Bioinf Comput Biol* **1**:1–20, 2003.
18. Daly M, *et al.*, High-resolution haplotype structure in the human genome, *Nat Genet* **29**:229–32, 2001.
19. Hinds D, *et al.*, Whole genome patterns of common DNA variation in diverse human populations, *Science* **307**:1072–1079, 2005.
20. Consortium IH, The international hapmap project, *Nature* **426**:789–796, 2003.

21. Qin Z, Nu T, Liu J, Partitioning-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms, *Am J Hum Genet* **71**: 1242–1247, 2002.
22. Stephens M, Donnelly P, A comparison of bayesian methods for haplotype reconstruction from population genotype data, *Am J Hum Genet* **73**:1162–9, 2003.
23. Halperin E, Eskin E, Haplotype reconstruction from genotype data using imperfect phylogeny, *Bioinformatics* **20**:1842–9, 2004.

**Eleazar Eskin** received his Ph.D. in Computer Science at Columbia University in October 2002. After graduation, he was a post-doctoral researcher at Hebrew University in Jerusalem, Israel. He is currently an Assistant Professor in Residence in the Department of Computer Science at the University of California, San Diego and affiliated with the California Institute of Telecommunications and Information Technology (Calit2).

**Roded Sharan** is a Senior Lecturer at the School of Computer Science, Tel Aviv University (since 2005), working in the areas of bioinformatics and systems biology. Prior to joining Tel Aviv University he was a post-doc at the University of California, Berkeley. He has a Ph.D. in Computer Science from Tel Aviv University, Israel.



**Dr. Eran Halperin** is a senior research scientist in the International Computer Science Institute (ICSI), in Berkeley, California. His research is focused on statistical and computational approaches in biology, particularly genetics, and applications to disease association studies. He received his M.Sc. and Ph.D. from the Computer Science department at Tel-Aviv University in Israel. He is a former postdoc at ICSI and UC Berkeley, and has held research positions at Princeton University and Compugen LTD, a bioinformatics company.