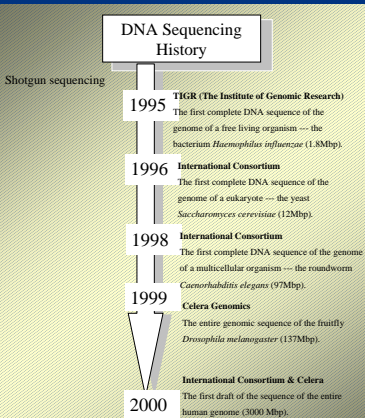


A dynamic programming algorithm for haplotype block partitioning

**K Zhang, MH Deng, T Chen,
MS Waterman, FZ Sun**

Center for Computational and Experimental Genomics
University of Southern California

Sequencing History



Human Variations

Different kinds of human variations:

- Substitutions
- Insertions/Deletions
- Duplications
- Rearrangements

Substitutions of a single nucleotide represent 10% to 50% of human variations.

SNP = single nucleotide polymorphism

SNPs in the human genome about 1 in 600bps.

Haplotype

SNP1

SNP2

C/T

G/T

There are four **haplotypes** possible:

C	_____	G
C	_____	T
T	_____	G
T	_____	T

In general for n SNPs, there are 2^n possible haplotypes.

Genotype

Since we are diploid (have two chromosomes), things are more complex. Genotype of an individual:

———— site 1 ———— site 2 ———— site 3 ————
 C/T G/T A/A

It tells us that the individual has different base pairs on the maternal and paternal chromosome at site 1 and 2, the same base pairs at site 3.

Haplotypes of an individual may be:

maternal ———— C ———— T ———— A ————
paternal ———— T ———— G ———— A ————

Our Data are from

Block of Limited Haplotype Diversity Revealed
by High-resolution Scanning of Human
Chromosome 21.

Patil N., Berno A.J., Hinds D.A., et al.

Science 294: 1719-1722 (23 Nov 2001).

Summary of the Data

- Sample 24 ethnically diverse individuals
- Separate the 2 copies of chromosome 21 using rodent-human somatic cell hybrid technology
- 20 independent copies of chromosome 21 analyzed
- 32.4×10^6 bases with 21.7×10^6 bases of unique sequence
- Essentially resequence using 3.4×10^9 oligonucleotide on 160 wafers
- Identity 35,989 SNPs
- Data at NCBI, dbSNP databases

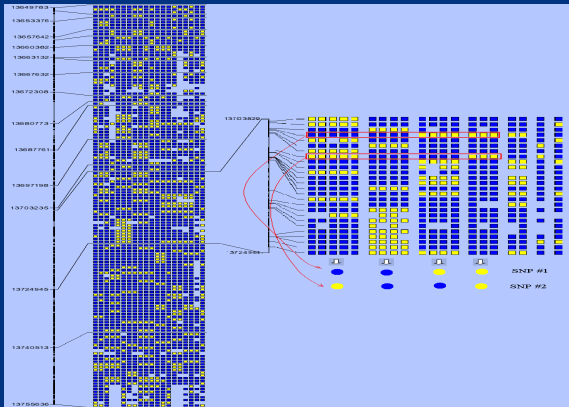
Haplotype Block Partitioning

- Objective:

To partition the haplotypes into blocks with minimum total number of SNPs required to account for most of the haplotype information in each block.

- Consecutive SNPs of size one or larger is a **block** only if the haplotypes represented more than once (**common haplotypes**) are more than a fraction (e.g., 80%).
- **Representative SNPs** in a block are SNPs that can distinguish at least a certain percentage (e.g., 80%) of haplotypes.

An example of Haplotype Blocks



The Greedy Algorithm

- Finding all blocks with corresponding number of representative SNPs with a given percentage.
- Calculating the ratio with the total number of SNPs in the block to the number of representative SNPs.
- Selecting block with maximum ratio and discarding all other blocks with overlap with this block.
- Previous process is repeated in the remaining blocks until the blocks with no gaps and with every SNP assigned to a block.
- The algorithm can not guarantee to minimum number of representative SNPs.

The Dynamic Programming Algorithm

Develop a dynamic programming algorithm to partition the haplotypes into blocks to minimize the number of SNPs required to account for most of the haplotype information in each block.

Features of the program:

- Any measure of haplotype information can be used
- Guaranteed minimum number of representative SNPs
- Relatively fast

Mathematical formulation

- The haplotypes are divided into blocks B_1, B_2, \dots, B_I .
- Let $A(B_i)$ be the minimum number of SNPs required to account for at least α percent of the haplotype information in block B_i .

Uniquely distinguish α percent of the unambiguous haplotypes OR α percent of block haplotype diversity

- **Objective:** Minimize the total number of SNPs required,

$$\sum_{i=1}^I A(B_i)$$

The dynamic programming algorithm

- Let $block(r_i, r_{i+1}, \dots, r_j) = 1$ if the SNPs r_i, r_{i+1}, \dots, r_j form a block satisfying certain conditions, and 0 otherwise.
- S_j : number of representative SNPs for the optimal block partition of the first j SNPs.

$$S_0 = 0$$

$$S_j = \min\{S_{i-1} + A(r_i, r_{i+1}, \dots, r_j), \\ \text{if } block(r_i, r_{i+1}, \dots, r_j) = 1\}$$

The dynamic programming algorithm

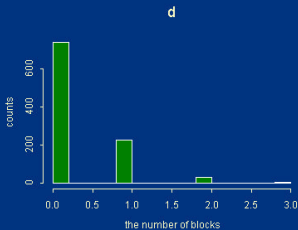
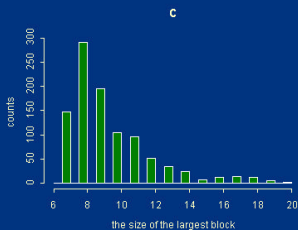
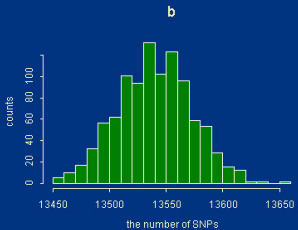
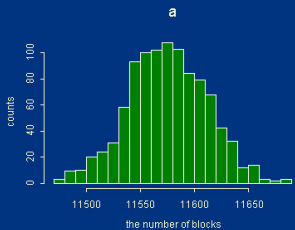
- First use the above dynamic programming algorithm to find S_n .
- Trace back to find the optimal block partition and the representative SNPs for the haplotypes.
- Finding the number of SNPs in a block is NP-hard. We use the enumeration method in this application.

Results

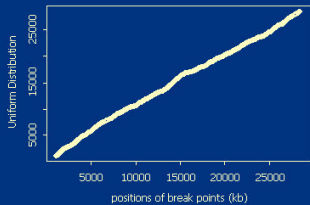
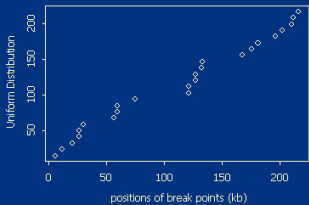
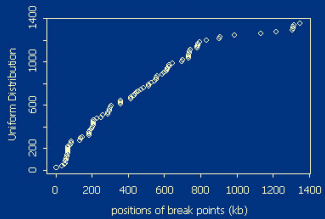
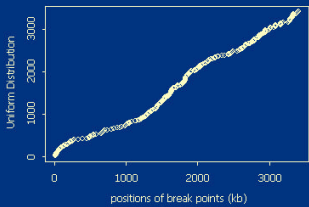
Block partition using dynamic programming vs greedy algorithm ($\alpha = 80\%$):

	number of SNPs	number of blocks	size of largest block	% blocks with > 10 SNPs
dynamic	3,582	2,575	128	28.8
greedy	4,536	4,135	114	14.2
difference	-21.0%	-37.7%	12.3%	102.8%

Testing the significance of the results using permutation tests



Homogeneity of the block break points

a**b****c****d**

Summary

- Develop a dynamic programming algorithm for haplotype block partition to minimize the number of representative SNPs.
- Apply the algorithm to the Chromosome 21 data. Test the statistical significance of the block partition results.
- Regions of biological interest can be identified based on the block break points.

Perspectives

- Faster algorithms to find the number of representative SNPs in a block may be needed for large scale problems
- How many chromosomes are needed to capture the common haplotype structures of the general populations?
- What are the biological reasons for the observed haplotype structure?
- For association studies, how much information can be lost (or gained) by using only the representative SNPs instead of all the SNPs?

Acknowledgements

- Thank Perlegen for making the data available and their excellent ideas on formulating the problem.
- USC computational biology group for many interesting ideas and discussions.
- Supported by NIH and NSF.