The human genomic SNP melting map project

Eivind Tøstesen and Eivind Hovig
Institute for Medical Informatics, Oslo University Hospital, Norway

NBF 2010

Introduction

The human genomic SNP melting map project aims to compute the variation in the thermodynamic DNA melting profiles for each SNP. Although a SNP is very localized, it has long-range effects on the double-helical stability in domains up to a few kilobases. It is a **High Performance Computing** challenge to predict these effects.

- The dotted curve in Fig. 1 is the human genomic melting map (MM).

- The coloured curve fragments in Fig. 1 are the human genomic SNP melting map (SNP-MM). Each colour shows the deviation from the MM due to one SNP.

The aim is to compute the SNP-MM for the whole genome.

HPC budget

In 2006, the MM was computed [ref. 1]. It took T_{hg}=22 CPU days on an HP Superdome (64 x Itanium 2 processors, 1.5 Ghz, 6 MB Cache). We may assume that the whole MM calculation was on the petaflop scale (=10^{15} floating point operations).

With the technology of 2006, the computing times for chromosome *j* scales as:

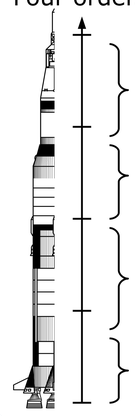
MM: $T_j = O(N_j)$

SNP-MM: $T_j = O(N_j \cdot S_j)$

where N_j is sequence length and S_j the number of SNPs. The whole chromosome must be processed for each SNP, effectively making the SNP-MM computation scale **quadratically**.
The total SNP-MM computing time is estimated to $5 \cdot 10^4 \cdot T_{hg}$, i.e. $2 \cdot 10^4$ years or 200 exaflop.

Exact approach

Four orders of magnitude **speedup** to reach petascale nirvana:



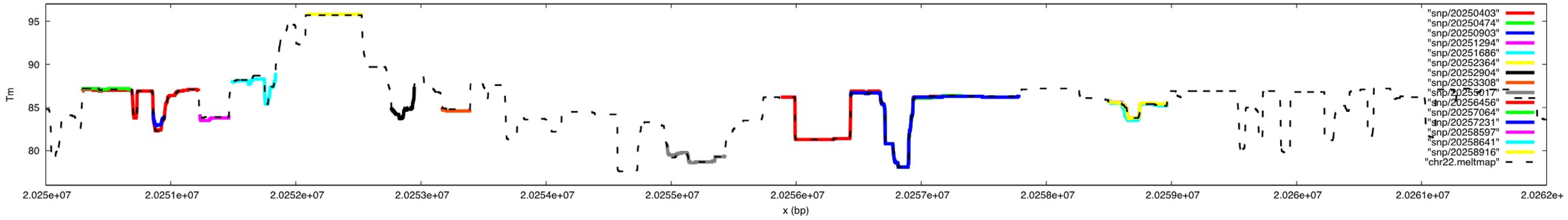
Reusing 20 GB intermediate calculations

bisection instead of looping over all temperatures

Parallelism per SNP or per chromosome

Moore's law

Figure 1: A 12 kb region In chr22



Approx approach

The **Alexandrian speedup**: For each SNP, consider only a sequence window of 12 kb (as in Fig. 1): 4 kb to capture the SNP effect and 4 kb flanking regions to dampen the DNA molecule's end effects. Compute each window twice (the two SNP variants) and compare.

The approximate SNP-MM computation scales as $T_j = O(S_j)$.

The estimated time becomes $8 \cdot T_{hg}$, i.e. 6 months with 2006 technology.

Problem: What size of flanking regions is necessary to minimize the approximation errors?

Discussion

The effect of multiple SNPs may not be additive. In that case, additional computation for overlapping SNPs (Fig. 1) of their combined effects could be required to predict individual genomes.

Further algorithmic speedup is possible. It may turn out that the exact SNP-MM also has time complexity $O(S_j)$.

A nearest neighbor Tm calculation may predict the rough stability changes for many SNPs with low cost, but not the extents of the changed regions.

References

[1] Liu F, Tøstesen E, Sundet JK, Jenssen TK, Bock C, Jerstad GI, Thilly WG and Hovig E. (2007) The human genomic melting map. PLoS Comput Biol 3(5): e93. doi:10.1371/journal.pcbi.0030093

[2] http://meltmap.uio.no