Short Specialist Review

Algorithmic improvements in gene mapping

Gonçalo R. Abecasis and Yu Zhao
University of Michigan, Ann Arbor, MI, USA

1. Introduction

The analysis of human gene mapping data has generated many challenging computational problems. The challenges arise because in most gene-mapping studies the DNA sequence of each individual is only measured imperfectly. For some individuals, these measurements are the result of genotyping assays at specific loci or chromosomal regions. For other individuals, there might be even greater uncertainty: their DNA sequence might only be measured indirectly through information obtained on the genotypes of their relatives. In either situation, there can be a very large number of DNA sequences compatible with observed data, and identifying the most likely DNA sequence configuration(s) might require many individuals to be considered jointly.

2. Analysis of human pedigrees

To survey algorithmic challenges in gene mapping, we will focus on the analysis of pedigree data. These data are often used in linkage studies of discrete or quantitative traits, in the construction of genetic linkage maps (*see* Article 54, **Sex-specific maps and consequences for linkage mapping**, Volume 1), in quality assessments for genotyping data, or to identify individual haplotypes. Maximum likelihood can solve these and other problems related to the analysis of pedigree data, and many algorithms have been developed to calculate likelihoods for human pedigrees. Briefly, the likelihood of interest can be written as:

$$L(\text{data}) = \sum_{G_l} \ldots \sum_{G_n} \prod_i P(X_i|G_i) \prod_{\text{founder}} P(G_{\text{founder}}) \prod_{\{o,f,m\}} P(G_o|G_f, G_m) \quad (1)$$

This likelihood involves a nested summation over the set of possible haplo-genotypes, G_i, for each individual. The likelihood of each possible configuration is a product with factors denoting (1) the probability of observed phenotypes conditional on each individual haplo-genotype, $P(X_i|G_i)$; (2) the probability of founder haplo-genotypes, $P(G_{\text{founder}})$; and (3) the probability of offspring haplo-genotypes conditional on parental haplo-genotypes, $P(G_o|G_f, G_m)$, calculated for all parent-offspring triples.

Direct evaluation of this nested sum is only possible in the simplest of cases, involving a very small number of loci and individuals. The number of summation terms to be evaluated increases exponentially with both the number of individuals in the pedigree (which add extra levels to the nested sum) and the number of markers being considered (which increase the number of possible haplo-genotypes for each individual).

In early gene-mapping studies, investigators painstakingly evaluated likelihoods for each pedigree examined, using careful algebra to factor the calculation and identify repeated terms. Gene mapping was a new field, laboratory methods were rudimentary allowing the use of only small amounts of data, and this laborious approach was adequate.

3. The Elston–Stewart and related algorithms

Elston and Stewart (1971) developed the first general algorithm for rapid pedigree likelihood calculation. They showed that the likelihood could be updated gradually, one nuclear family at a time. Each update required iterating over possible genotypes for individuals in the nuclear family, resulting in a relatively small nested sum. Their strategy proved highly effective, and their algorithm is still the method of choice for the analysis of large, noninbred pedigrees. Their method was later implemented in LIPED, the first widely available automated software for pedigree likelihood calculation (Ott, 1976), which played a crucial role in enabling the gene-mapping revolution.

Many improvements to the basic algorithm have been proposed. For example, Cannings *et al*. (1978) showed how the method could – in theory – be applied to complex pedigrees, even with inbreeding. Lange and Boehnke (1983) showed that the likelihood could be updated one individual, rather than one nuclear family at a time, and that different sequences of updates could produce dramatically different computing time and memory requirements. With these improved formulations, the complexity of calculating likelihoods for most noninbred pedigrees increases linearly with the number of individuals in a family, and likelihoods can be calculated for very large pedigrees, including hundreds of individuals. Another important enhancement was the development of algorithms for identifying sets of haplo-genotypes for each individual compatible with the observed genotypes for each family (Lange and Goradia, 1987).

In parallel to these algorithmic improvements, more sophisticated computer implementations of the Elston–Stewart algorithm were developed. The LINKAGE computer package (Lathrop *et al*., 1984; Lathrop *et al*., 1985) enabled geneticists to analyze multiple marker loci jointly. Together with the discovery of highly polymorphic VNTR and microsatellite markers, LINKAGE enabled the localization of genes for many Mendelian disorders through multilocus linkage analysis in relatively large pedigrees.

In the 1990s, further enhancements to the Elston–Stewart algorithm were discovered. Cottingham *et al*. (1993) used improved software engineering techniques, such as caching and replacement of floating point with integer operations, to speed up LINKAGE by about one order of magnitude. O'Connell and Weeks (1995) showed that combining alleles that do not appear in an individual's descendants in a single meta-allele could dramatically reduce the number of distinct haplo-genotypes and further speed up calculation.

Despite these enhancements, the complexity of likelihood calculations using the Elston–Stewart algorithm grows exponentially with the number of marker loci considered. State-of-the-art implementations of the Elston–Stewart algorithm in the VITESSE (O'Connell and Weeks, 1995) and FASTLINK (Cottingham *et al*., 1993) computer packages cannot handle more than 5–10 marker loci at a time. The ability of geneticists to rapidly collect data for hundreds of microsatellite markers and the interest in complex disease gene mapping using large collections of small pedigrees shifted the focus to a different collection of algorithms.

4. The Lander–Green and related algorithms

Lander and Green (1987) proposed a very different strategy for pedigree likelihood calculations. Their approach is based on the use of inheritance vectors, which summarize inheritance at specific genomic location. They showed that the probability of observed genotypic or phenotypic data can be calculated for any particular inheritance vector and that, in the absence of genetic interference, inheritance vectors form a Markov Chain along the chromosome. Using a Hidden Markov Model, they proposed an algorithm for the calculation of pedigree likelihoods whose complexity increased only linearly with the number of markers. The algorithm is suitable for very large numbers of markers, but limited to relatively small pedigrees because the number of possible inheritance vectors increases exponentially with pedigree size.

As with the Elston–Stewart algorithm, many enhancements were later discovered, and progressively more powerful computer implementations contributed to the success of countless gene-mapping studies. One important enhancement resulted from the observation that there are many redundancies within inheritance vector space so that inheritance vectors can be grouped to speed up calculation. Over the years, progressively more sophisticated strategies were developed for identifying these redundancies, first focusing on symmetries resulting from the transmission of alleles from single founders (Kruglyak *et al*., 1996), then founder couples (Gudbjartsson *et al*., 2000), and later from other individuals in the pedigree (Markianos *et al*., 2001; Abecasis *et al*., 2002). Another important series of improvements focused on the manipulation of transition matrices, used for the calculation of conditional inheritance vector distributions at neighboring locations, a key step in the Markov Chain. Two distinct approaches have proved very successful at speeding up this step of the calculation: either a divide-and-conquer algorithm (Idury and Elston, 1997) or Fast Fourier Transform (Kruglyak and Lander, 1998) can reduce the computational cost of generating these conditional distributions by several orders of magnitude.

Popular implementations of the Lander–Green algorithm include the computer packages GENEHUNTER (Kruglyak *et al*., 1996; Markianos *et al*., 2001), ALLEGRO (Gudbjartsson *et al*., 2000), and MERLIN (Abecasis *et al*., 2002). All these packages can handle very large numbers of markers and allow the estimation of individual haplotypes or the analysis of quantitative and discrete traits, providing parametric and nonparametric linkage tests. They also provide more specialized algorithms including, for example, algorithms that estimate information content along the genome. Relative information content can highlight areas where genotyping additional markers would provide the greatest information gain (*see* Article 53, **Information content in gene mapping**, Volume 1). In addition to these standard features, the newer packages can generate simulated datasets (commonly used for calculating empirical significance levels), carry out more accurate linkage tests (Kong and Cox, 1997; Sham *et al*., 2002), and even identify likely genotyping errors (Abecasis *et al*., 2004).

Although current implementations of the Lander–Green algorithm can comfortably handle hundreds and even thousands of genetic markers, advances in laboratory technology are already highlighting a need for even more powerful methods. The shift to SNP (single-nucleotide polymorphism) markers and very large scale genotyping has generated datasets with hundreds of thousands of markers measured per individual, with substantial amounts of linkage disequilibrium between neighboring markers (*see* Article 50, **Gene mapping and the transition from STRPs to SNPs**, Volume 1). It is likely that further enhancements to gene-mapping algorithms will be forthcoming to allow the analysis of these new datasets.

5. Markov-chain Monte-Carlo algorithms

While the Elston–Stewart and related algorithms can handle a small number of markers in very large noninbred pedigrees and the Lander–Green and related algorithms can handle very large numbers of markers in small pedigrees, neither approach can handle a large number of markers in a large pedigree. Very large pedigrees arise in many interesting settings, most often in the study of isolated populations (*see* Article 51, **Choices in gene mapping: populations and family structures**, Volume 1). It is often desirable to analyze multiple genetic markers in these pedigrees to clarify inheritance patterns when genotype data are not available for individuals in the early generations. The analysis of these most challenging datasets has motivated the development of Monte-Carlo-based methods, which try to identify the most important terms in the pedigree likelihood but avoid summing over all possible terms.

A variety of Monte-Carlo approaches have been employed successfully in linkage analysis, including Simulated Annealing (Sobel and Lange, 1996, implemented in the SIMWALK2 computer program), the Gibbs sampler (Heath, 1997, implemented in the LOKI computer program), and Sequential Imputation (Irwin *et al*., 1994). In addition to the ability to handle very large datasets, these software packages often provide capabilities not currently available in packages based on the Elston–Stewart or Lander–Green algorithms. For example, LOKI (Heath, 1997) can model the contributions of multiple susceptibility loci simultaneously and

SIMWALK2 (Sobel and Lange, 1996; Sobel *et al*., 2002) can model genotyping error explicitly.

6. Outlook for the future

While this is an incomplete account of all the algorithms developed for the linkage analysis of human pedigrees, we have attempted to emphasize those algorithms and developments that entered widespread use through the availability of easy-to-use computer programs. We have certainly missed some packages and ideas that deserve credit, as well as some research paths and strategies that were tried along the way but never become popular in practice. Currently, one promising avenue appears to be the use of Bayesian Networks (Jensen, 1996). These allow complex likelihoods to be evaluated gradually, and provide for a more flexible updating scheme than the Lander–Green or Elston–Stewart algorithms, which conduct updates considering either all individuals (for one locus) or all loci (for one or more individuals) at a time.

The past 20–30 years have produced many algorithmic advances in the analysis of human pedigrees, and these have enabled geneticists to extract the full benefits of new laboratory methods that allow the collection of increasing amounts of genetic information on increasing samples of individuals. Whereas initial methods focused on the analysis of single genetic markers and simple Mendelian traits, more modern methods can analyze very large numbers of genetic markers and individuals and have led to some promising results in the analysis of even complex traits such as diabetes, asthma, and psychiatric disorders. It is tempting to speculate that, with the increasing emphasis on genetic association studies and fine-mapping data (Cardon and Abecasis, 2003), the next decade will produce similar advances in algorithms for the estimation and analysis of haplotypes . . . , but we will leave that story for the 2nd edition!

References

Abecasis GR, Burt RA, Hall D, Bochum S, Doheny KF, Lundy SL, Torrington M, Roos JL, Gogos JA and Karayiorgou M (2004) Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *American Journal of Human Genetics*, **74**, 403–417.

Abecasis GR, Cherny SS, Cookson WO and Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, **30**, 97–101.

Cannings C, Thompson EA and Skolnick MH (1978) Probability functions on complex pedigrees. *Advances in Applied Probability*, **1**, 26–61.

Cardon LR and Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. *Trends in Genetics*, **19**, 135–140.

Cottingham RW Jr, Idury RM and Schaffer AA (1993) Faster sequential genetic linkage computations. *American Journal of Human Genetics*, **53**, 252–263.

Elston RC and Stewart J (1971) A general model for the genetic analysis of pedigree data. *Human Heredity*, **21**, 523–542.

Gudbjartsson DF, Jonasson K, Frigge ML and Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics*, **25**, 12–13.

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, **61**, 748–760.

Idury RM and Elston RC (1997) A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. *Human Heredity*, **47**, 197–202.

Irwin M, Cox NJ and Kong A (1994) Sequential imputation for multilocus linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 11684–11688.

Jensen FV (1996) *An Introduction to Bayesian Networks*, University College Press: London.

Kong A and Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics*, **61**, 1179–1188.

Kruglyak L, Daly MJ, ReeveDaly MP and Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, **58**, 1347–1363.

Kruglyak L and Lander ES (1998) Faster multipoint linkage analysis using Fourier transforms. *Journal of Computational Biology*, **5**, 1–7.

Lander ES and Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 2363–2367.

Lange K and Boehnke M (1983) Extensions to pedigree analysis. V. optimal calculation of Mendelian likelihoods. *Human Heredity*, **33**, 291–301.

Lange K and Goradia TM (1987) An algorithm for automatic genotype elimination. *American Journal of Human Genetics*, **40**, 250–256.

Lathrop GM, Lalouel J, Julier C and Ott J (1984) Strategies for multilocus linkage in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **81**, 3443–3446.

Lathrop GM, Lalouel JM, Julier C and Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics*, **37**, 482–498.

Markianos K, Daly MJ and Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of inheritance space. *American Journal of Human Genetics*, **68**, 963–977.

O'Connell JR and Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics*, **11**, 402–408.

Ott J (1976) A computer program for general linkage analysis of human pedigrees. *American Journal of Human Genetics*, **26**, 588–597.

Sham PC, Purcell S, Cherny SS and Abecasis GR (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics*, **71**, 238–253.

Sobel E and Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, **58**, 1323–1337.

Sobel E, Papp JC and Lange K (2002) Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics*, **70**, 496–508.