severe portal hypertension. And there was the added risk of bleeding after the removal of the venous catheter. Considering these and many other risks perhaps the patient was fortunate to undergo the procedure with so few complications. On the other hand, she suffers from a fatal illness which is not responding well to conventional therapy. It is only necessary to think back to the early days of cardiac transplantation to remember that most major advances in high technology medicine have been fraught with dangers and uncertainties. The ethical questions arising from the difficult decision of when to move from experimental animals to patients are very complex, and must be worked through at leisure by knowledgeable ethics committees, and well-informed patients and their families and those who care for them; from what we can glean these issues seem to have been handled with great care in this case.

Despite all these if's and but's, the work described by Grossman and her colleagues represents a genuine step forward in the slow road to successful somatic gene therapy. It suggests that, ultimately, it will be possible to correct genetic diseases which are expressed primarily in liver cells, which include, in addition to familial hypercholesterolaemia, α1-antitrypsin deficiency and phenylketonuria. In the longer term, however, it will be important to evolve less traumatic approaches to introducing genes into liver cells; recent successes in targetting vectors to the liver suggest that this may be possible before long[6]. And there are still doubts about the longevity and levels of expression of human genes introduced into cells using vectors of this type.

Gene therapy was never going to be easy; the work of Grossman and her team, together with apparent successes in short-term gene replacment therapy in peripheral blood lymphocytes[11], should be of great encouragement to workers in the field. □

Computing the genetic map

G. Mark Lathrop

INSERM U.358, Institut de Génétique Moléculaire, 27 rue Juliette Dodu 75010, Paris, France

The pace of genetic mapping is continuing to accelerate, and with this comes the need for new computing software for data analysis and distribution of map information. Two papers in this issue of *Nature Genetics* describe important advances in genetic mapping software that respond to some of these pressing computing needs.

In October of 1992, the NIH/CEPH genetic map was published[1]. This consisted of a compilation of 1,416 loci from the CEPH database, largely restriction fragment length polymorphism (RFLP) markers that had been contributed by many different laboratories over the previous ten years. The same month saw the publication of a 2nd generation map, containing more than 800 CA-repeat microsatellite markers from the Généthon group[2]. An update of the Généthon map with more than 2,000 CA-repeat markers will be published soon[3], and other groups are producing similar maps based on microsatellite makers with CA or other repeat elements[4,5]. A prediction (easy to believe) is that 1995 will see the publication of a map containing 5,000–10,000 highly polymorphic, and easily characterized PCR-based markers. This will be a important resource for the study of genetic diseases, and a significant advance for the genome programme. But will it be the final stage of the human genetic map? Probably not, as many more polymorphic markers will be developed for the studies of disease association, linkage disequilibrium and variation within genes.

Existing computational and database tools for constructing and distributing maps have been strained by the rapid growth of data. In this issue, Matise *et al.*[4] introduce a new software package, MultiMap, for automated map construction. The software can be obtained through electronic mail, and it is easy to install and operate in any laboratory having access to minimum workstation hardware. It allows rapid construction of genetic maps, and verification of genotype data with little human intervention.

In a second paper, Buetow *et al.*[5] from the Cooperative Human Linkage Center (CHLC) describe a similar, but slightly less automated system for map construction. Most importantly they have taken a first step towards on-line network access of genetic map information, and have created a server for linkage computations which accepts submission of genotype data on CEPH families from outside laboratories and returns preliminary mapping information via electronic mail. An extension of the server to provide full integration of loci into the map and error analysis is promised for the near future. Large integrated genome maps (that is, maps combining data from different laboratories) constructed from genotypes from the CEPH database illustrate both software systems, and the CHLC map contains 214 previously unpublished microsatellite markers.

One interesting feature of these packages is the manner in which they extend and build upon statistical methods and software that have been verified for earlier RFLP-based maps. The construction of a genetic map relies on an underlying procedure to calculate maximum likelihood estimates of recombination rates. Both these packages use the CRI-MAP program[6] as the analytic engine for their calculations. The integration of pre-existing programs into new and more versatile packages represent an important simplification in software development. Similarly, another group has recently implemented new algorithms for rapid calculation of likelihoods in general pedigrees based on adaption of the LINKAGE programs[7].

CRI-MAP, and most other programs for genetic map construction, obtain recombination estimates by the maximum likelihood method, which is a standard technique in

Lessons of map construction

The first task involved in map construction is to assign new markers to chromosomes by calculation of lod scores with loci from existing maps. In the past, it was necessary to assign some of the markers by physical methods, but now genetic maps are sufficiently dense that they can always be assigned by linkage. Once markers are assigned to a chromosome, the next step is to determine their probable order. This is achieved by comparing the maximum likelihood values under the different orders (for simplicity, the maximum is often called "the likelihood of the order"). The best-supported order — that with the highest likelihood — is chosen as the order estimate.

Because we are dealing with statistical inference, it is impossible to conclude that the best-supported order is true. However, its likelihood can be compared to those for alternative orders to determine which of the latter can be conclusively rejected. This is achieved through calculation of an odds, which is the ratio of their likelihoods. The odds are usually written in a manner that emphasizes the concept of a ratio, such as 10:1, 100:1, 1000:1 and so on.

In maximum likelihood theory, the odds can be interpreted as a likelihood ratio test (LRT) statistic. For large values of the LRT (large odds in favour of the maximum likelihood order) the alternative order is said to be rejected. Although the LRT is a classical procedure in statistical hypothesis testing, its application as a test of gene order poses technical problems which cannot be discussed here. Due to these problems, it is impossible to calculate a theoretical significance level or p-value for the LRT in this setting. Thus the usual criterion for the rejection of alternative orders, odds of 1000:1 or greater, is somewhat of an arbitrary choice. It was initially adopted by analogy with a lod score of 3 (odds ratio of 1000:1) used as a critical value to detect linkage between two loci. However, much practical experience has shown that the 1000:1 criterion is very conservative, and leads to reliable inference of gene order.

Ideally, the likelihoods of all orders should be compared for construction of the map. In practice, this is impossible because of their number. For a map containing n loci, the number orders is by $n!/2=n\times(n-1)\times(n-2)...\times1/2$, where n! is the number of permutations of n items in a list, and the division by 2 corrects for the fact that mirror orders are equivalent. For example, the number of orders for 15 loci is 653,837,184,000; to obtain the number for 20 loci, it would be necessary to multiply this figure by 1,860,480. The maps of the smallest autosomal chromosomes now contain substantially more than 20 markers, and some published here contain more than 200. It is obvious that all orders cannot be examined even with the most rapid of computers. Fortunately, reliable heuristic algorithms can be designed that search a subset of orders to find the best-supported map and likely alternatives as illustrated by the MultiMap[4] and CHLC[5] approaches. □

applied statistics. In genetic mapping studies, the likelihood is the probability density (which is often shortened to "the probability") for the observed transmission of alleles from one generation to the next. In the absence of interference, the probability density is a function only of the recombination rates between adjacent markers. The maximum likelihood estimates of the recombination rates are the values of these parameters that maximize the likelihood, assuming a fixed order for the marker loci. In general, the problem of finding maximum likelihood estimates for recombination rates in multilocus analysis is difficult, but the fixed structure of the CEPH reference families (children, parents and possible grandparents) leads to simplifications that have been exploited in CRI-MAP to make it an efficient estimation program.

MultiMap and the CHLC software use the likelihood engine provided by CRI-MAP to implement algorithms for map construction. The algorithms themselves are based on previous experience with RFLP-based maps, but the new implementations provide much easier use. To understand these approaches, it may be helpful to review the steps involved in map construction (see Box).

The algorithms contained in MultiMap and the CHLC software are based on the idea that a well-supported map is unlikely to be widely perturbed by the introduction of a new marker. These programs construct a map by stepwise procedures, in which two markers are selected to form an initial map, and new markers are added based on their polymorphic information content (PIC). New markers are tested in all intervals, and the marker is retained in the map only if it can be placed in a single interval with odds superior to 1000:1, or with other criteria chosen by the user; otherwise, a series of possible positions is provided. In the MultiMap program, the construction of the map is followed by testing of alternative orders based on permutations of 2, 3 or more neighbouring loci, to determine if an order with higher support can be found. It is usual to distinguish a framework map consisting of a subset of loci for which the best- supported order has odds of 1000:1 greater than any alternative order.

Another area where the two software systems provide new features is the integration of error detection. Undetected genotype errors are one of the principal difficulties in attempting to resolve locus order in a genetic map. Statistical techniques can be used to detect a large proportion of these problems. Usually this requires two separate procedures. The first involves the identification of families that contain several apparent recombinants between closely linked loci. It is extremely important that these clusters should be identified and the data verified, as a single genotype error in the parental or grandparental generation may lead to inference of an incorrect parental phase, and the erroneous scoring of non-recombinant meioses as recombinant. These corrections require only analysis of recombination events between pairs of loci, and is independent of the locus order; it is undertaken automatically by the CHLC software. Errors in the offspring generation usually remain undetected until after a locus order has been inferred and apparent double or triple recombinants are found. It is very important to detect and verify apparent multiple recombinants, since such errors can be critical in the inference of locus order.

The two mapping systems considered here carry the error detection procedures a step further. The reliability of each locus is evaluated by its effect on the length of a map, calculated from the change in the estimated length when a marker is removed from the map. Genotype errors are responsible for inflation of the estimated length due to the presence of erroneous recombination events. Once unreliable loci are identified, they can be removed or assigned low priority for entry into a framework map.

Given the availability of this new

software, what future directions need to be explored? Several developments are likely to be important. One is the integration of the genetic and physical maps. The MultiMap program already takes one step in this direction, by allowing for analysis of radiation hybrid data, and incorporation of physical locations to constrain locus order. More significant is the role of a high resolution microsatellite-based genetic map as a source of ordered STSs for anchoring physical contigs, as illustrated in recent work[8]; new database and analysis tools are being created to integrate the two types of information. Efficient access to both genetic and physical maps are needed for disease linkage analysis. Computational servers available through electronic mail for rapid calculations of linkage in disease pedigrees would be a service for many laboratories, and some such systems are likely to be available in the near future. We can hope that they will link automatically to the mapping database to draw upon the latest information on locus orders and genetic distances between markers. Finally, the distribution of the CEPH genotype data for high resolution microsatellite maps should spur new investigations of interference, variation of recombination by sex, and other parameters of the human genetic map. □

1. NIH/CEPH Collaborative Mapping Group. *Science* **258**, 67–86 (1992).
2. Weissenbach, J. *et al. Nature* **359**, 794–801 (1992).
3. Weissenbach, J. *et al. Nature Genet.* (in the press).
4. Matise, T.C., Perlin, M. & Chakravarti, A. *Nature Genet.* **6**, 384–390 (1994).
5. Buetow, K.H. *et al. Nature Genet.* **6**, 391–393 (1994).
6. Lander, E. & Green, P. *Proc. natn. Acad. Sci. U.S.A.* **84**, 2363–2367 (1987).
7. Cottingham, R.W., Idury, R.M. & Schffer, A.A. *Am. J. hum. Genet.* **53**, 252–263 (1993).
8. Cohen, D., Chumakov, I. & Weissenbach, J. *Nature* **366**, 698–701 (1993).

Cystinuria defect expresses itself

Ernest M. Wright

Ernest M. Wright is Professor and Chair of Physiology at UCLA School of Medicine, Los Angeles, California 90024, USA

Cystinuria, a common inherited disorder, affects about 1 in 15,000 people in the United States and 1 in 2,000 in Europe. Cystinuria is caused by a renal cystine transport defect; each day the kidney filters 180 litres of plasma, and excretes 1.5 litres of amino-acid-free urine. A defect in the cystine reabsorptive mechanism results in the excretion of all the filtered cystine in the urine. This defect and the low cystine solubility in concentrated urine, favours the formation of cystine stones. Kidney stones are usually benign, although they do in some cases cause obstruction and intractable pain.

Although cystinuria (first described by Sir Archibald Garrod, in 1908) and other genetic defects of renal amino acid transporters have been well known for more than half a century, the identities of the membrane proteins and the nature of the defects have remained a puzzle, until recently. The problem is that transporters are rare (comprising less than 0.01% of membrane proteins), greasy molecules that have resisted identification and purification. A breakthrough came seven years ago with expression cloning in *Xenopus* oocytes, where membrane transport proteins were cloned using transport assays[1]. Using this method, three groups reported the cloning of a renal membrane protein that induces the expression of cystine transport in *Xenopus* oocytes (variously termed NAA-Tr (ref. 2); D2 (ref. 3) and rBAT (ref. 4). The 90 kDa type II glycoprotein stimulated cystine, dibasic and neutral amino acid uptake into oocytes with kinetics similar to the renal brush border transporter. This led to the speculation that a defect in the human form of rBAT caused the disease[5,6].

Two papers testing this hypothesis in different ways appear in this issue of *Nature Genetics*[7,8]. Kastner and his colleagues at the National Institutes of Health, in collaboration with others in Israel and France, have provided suggestive evidence from linkage analysis that *rBAT* is indeed the gene causing the disease. The earlier mapping of *rBAT* to chromosome 2 (ref. 5) provided the stimulus to examine linkage between chromosome 2 markers and cystinuria in 17 affected families. Their success was facilitated by the choice of patients—Libyan Jewish families, in which incidence of the disease is very high (1 in 2,500), and a high percentage of patients come from consanguinous marriages. Three chromosome 2p markers cosegrated with the disease with high (>3.00) lod scores. Given the nature of the renal transport defect and the properties of rBAT expressed in oocytes, this linkage implicated the *rBAT* gene as the primary candidate for cystinuria. Proof of this hypothesis is provided by Palacín and colleagues[7] in Spain and Italy, using in part an approach pioneered by Turk *et al.*[9] for defining the molecular basis of the intestinal disease, glucose-galactose malabsorption.

Palacín's starting point was human *rBAT* cDNA[8], and their strategy was to isolate mRNA from patients with cystinuria, amplify overlapping segments of *rBAT* cDNA and screen for mutations using SSCP analysis. Six specific mutations, including homozygotes and compound heterozygotes, were found that segregated with the cystinuria phenotype. A diversity of mutations is expected with autosomal recessive defects, as indicated by the more than 350 mutations so far known for the cystic fibrosis gene. The most common mutation in Spanish cystinuria patients changed methionine 467 to threonine. Expressing this Met467Thr mutant protein in oocytes produced only 20% of wild type L-cystine, L-arginine and L-leucine transport activity. This leads to the convincing conclusion that mutations in *rBAT* cause the transport defect underlying cystinuria.

But what exactly is the role of rBAT in cystine transport? The question arises because the overwhelming majority (>120) of transport proteins cloned from bacteria, plants and animals belong to a family of membrane proteins where the molecule threads its way across the lipid bilayer 12 times[10], whereas rBAT probably only spans the membrane once. Is rBAT a transport protein or a