Intelligent DNA-Based Molecular Diagnostics Using Linked Genetic Markers

Dhiraj K. Pathak[1], Eric P. Hoffman[2], and Mark W. Perlin[1]
[1] Department of Computer Science, Carnegie Mellon University
[2] Department of Molecular Genetics and Biochemistry, University of Pittsburgh

Abstract

This paper describes a knowledge-based system for molecular diagnostics, and its application to fully automated diagnosis of X-linked genetic disorders. Molecular diagnostic information is used in clinical practice for determining genetic risks, such as carrier determination and prenatal diagnosis. Initially, blood samples are obtained from related individuals, and PCR amplification is performed. Linkage-based molecular diagnosis then entails three data analysis steps. First, for every individual, the alleles (i.e., DNA composition) are determined at specified chromosomal locations. Second, the flow of genetic material among the individuals is established. Third, the probability that a given individual is either a carrier of the disease or affected by the disease is determined.

The current practice is to perform each of these three steps manually, which is costly, time consuming, labor-intensive, and error-prone. As such, the knowledge-intensive data analysis and interpretation supersede the actual experimentation effort as the major bottleneck in molecular diagnostics. By examining the human problem solving for the task, we have designed and implemented a prototype knowledge-based system capable of fully automating linkage-based molecular diagnostics in X-linked genetic disorders, including Duchenne Muscular Dystrophy (DMD). Our system uses knowledge-based interpretation of gel electrophoresis images to determine individual DNA marker labels, a constraint satisfaction search for consistent genetic flow among individuals, and a blackboard-style problem solver for risk assessment. We describe the system's successful diagnosis of DMD carrier and affected individuals from raw clinical data.

Introduction

The Human Genome Project is a world-wide effort whose completion will result in the localization, cloning, and sequencing of all 100,000 human genes (Watson 1990). Once a disease gene sequence is known, it can be used for molecular diagnosis, determining gene function, and, perhaps, for eventual gene therapy of the disease. This paper focuses on the first use: molecular diagnostics via the polymerase chain reaction (PCR) (Saiki *et al.* 1988). We do this by genetic linkage analysis (Ward *et al.* 1989) of multiple, highly polymorphic, PCR-based genetic markers that lie within or near the disease gene. By performing multiplex PCR on each individual in a family, intelligent analysis of the genetic flow patterns can determine the probability that an individual is affected by (or carries) the disease. Such genetic testing is crucial in effectively counselling at-risk individuals for family planning decisions. Further, our approach is generally applicable to any disease and its gene(s).

Current practice (e.g., multiplex PCR (Chamberlain *et al.* 1990), abundant mapped markers, automated DNA sequencers) automates much of the laboratory experimentation, and can generate vast amounts of genetic data quite rapidly. For successful linkage analysis, highly informative genetic markers are used (e.g., PIC >.70 (Botstein *et al.* 1980)), with many possible alleles for each marker. For ease of experimentation and allele detection, simple tandem repeat (STR) markers are used. STRs are readily amplified via PCR, and, since the alleles differ only by the number of tandem repeats (i.e., the size of DNA molecule), the alleles are easily detected by size separation on electrophoretic gels.

Ideally, a hemizygote (one chromosome having one allele at the marker) or homozygote (two chromosomes having the same allele at the marker) would generate one band on a gel. The location of this band on a routine or DNA sequencer gel corresponds to the size of the allele's DNA product. Similarly, a heterozygote (two chromosomes having different alleles at the marker) would generate precisely two bands on a gel that correspond to the two allele sizes. This determination of alleles for the genetic markers in a family is termed *genotyping*. However, with CA-repeat dinucleotide STR markers (Weber & May 1989) (which are highly abundant in the genome and generally used for genotyping), stutter bands occur that produce more complex patterns.

In practice, the requisite data acquisition through

PCR and sizing experimentation is just the first (and often least difficult) step in molecular diagnostics. The subsequent knowledge-intensive information processing is often the key bottleneck:

- Determining the alleles (i.e., the genotypes) from the complex stuttered banding patterns in sizing signal.
- Determining each individual's haplotypes (i.e., which alleles lie on which chromosomes) from the genotyping information.
- Assessing each individual's risk of disease (or carrier status) from this haplotyping information, and any other phenotypic information.

Each component step requires intelligent knowledge-based processing, and must provide recovery mechanisms to work in the face of incomplete or inconsistent data.

The major (and increasing) proportion of molecular diagnostics time is now devoted to these information processing tasks, for which there is little computational support. The state-of-the-art in software systems for automatic data acquisition (eg. (App 1993)) assumes interactive processing of the acquired data by the user. In contrast, our efforts are focused on automating the complex task of data interpretation after the data has been acquired.

In this paper we describe an intelligent system that frees the user entirely of the tasks of visually analyzing the acquired signals, establishing data consistency, and computing disease risks. The prototype system is described in Section 2 and its component modules are described in Sections 3, 4, and 5. A detailed example of its use is presented in Section 6 with specific application to the intelligent molecular diagnosis of Becker/Duchenne Muscular Dystrophy (B/DMD). Concluding observations and future work are discussed in Section 7.

An Intelligent System for Fully Automating Molecular Diagnosis

Figure 1 illustrates the system architecture consisting of three modules. The first module uses a rule-based system (Hayes-Roth 1985) to intelligently interpret sizing signals to find possible alleles for each locus. The second module uses constraint satisfaction (Bibel 1988) to reason with family constraints to infer possible genotypes from the possible alleles. The third module uses a blackboard problem-solver (Lesser *et al.* 1975; Engelmore & Morgan 1988) to automatically compute risks for possible genotypes.

B/DMD is a sex-linked skeleto-muscular disease of young males (Emery 1988) which occurs in 1/3500 live births. It is caused by a deleterious modification of the (extremely large) 1.5 megabase Dystrophin gene on the X chromosome (Koenig *et al.* 1987); the normal Dystrophin protein product acts at the muscle cell membrane to stabilize ion flow. Generally, a family clinically presents with an affected relative, and is interested in ascertaining which family members are carriers of, or are affected with, the disease. A genetic counselor records the family pedigree (i.e., inheritance graph structure), and determines useful phenotypic (affected, carrier, serum Creatine Kinase, etc.) information. Blood samples are then taken from informative individuals, on which PCR amplification of intragenic loci performed. In roughly half of families, the exon deletion analysis (Chamberlain *et al.* 1990) is equivocal, and STR-based linkage analysis is performed.

Molecular diagnostics in a family begins with the pedigree structure, phenotypic information, and each individual's gel sizing data from four multiplexed intragenic STR markers (Schwartz *et al.* 1992). Our molecular diagnosis system, shown in Figure 1, replicates the expertise of highly trained molecular geneticists at each of three central steps:

1. Allele determination by intelligent interpretation of gel sizing signals, which deduces allele sizes from complex stutter patterns. This machine vision task is done in our system by a rule-based interpretation module.
2. Determining the chromosomal flow of the B/DMD gene within the family to determine haplotypes. Our system does this by a constraint satisfaction process that sets the phase of alleles, assigning each chromosome a linear allele sequence.
3. Computing disease risk for individuals. Our system has multiple strategies for this, including using the haplotype as a disease gene signature for symbolic computation, and Bayesian blackboard problem solving for probabilistic computation.

The output is then visually presented to the user via an interactive interface that shows the pedigree. The system is implemented in Common LISP/CLOS, with a Macintosh user interface. Once the input resides in main memory, analysis proceeds in several seconds on a Quadra-class machine.

Knowledge-Based Interpretation of Sizing Signal

To determine the alleles at specified loci of every individual of an at-risk family a sizing experiment is performed. As illustrated in schematic form in Figure 2, the starting point is blood samples from individuals in the affected family. First, the DNA is extracted from the blood sample. Then, multiplex PCR is used to amplify the DNA at specified loci. The PCR product is labeled with fluorescent dye and electrophoresed on a polyacrylamide gel. The amount of migration as a function of time is the *sizing signal.* Several reference markers are added to the PCR products for calibration purposes.

Our system can automatically interpret two common types of sizing signals illustrated in Figure 3. The first

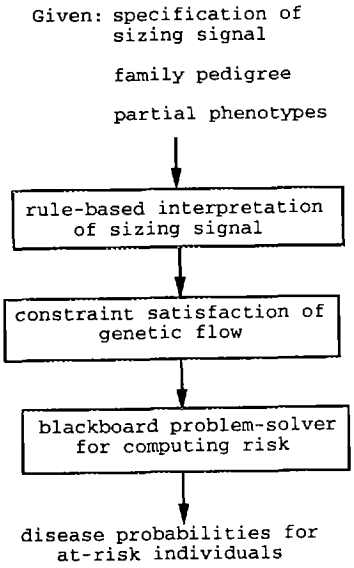

Figure 1: The overall architecture of the molecular diagnostics system.

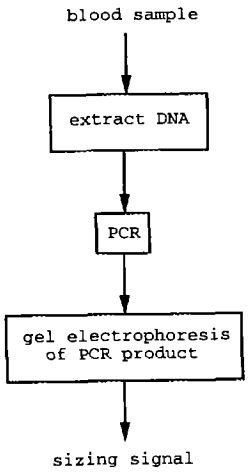

Figure 2: The acquisition of sizing data.

sizing signal, represented in Figure 3 (a), is obtained using a Dupont Genesis automated DNA sequencer and consists of a discrete sequence of intensity values for each lane on a gel. Usually multiple lanes, each with multiple PCR products and reference markers are run on the same gel. In this signal, the peaks are the significant features. s(k) denotes the sampling signal and k is the sampling unit. In the example shown, there are two peaks corresponding to two reference markers. These peaks are located at k1 and k2. Additionally, there are peaks corresponding to the alleles for each locus whose PCR products are included in the lane. Two such peaks are located at k3 and k4. (It is usual to refer to loci as genetic markers, or markers for short.) The second sizing signal, represented in Figure 3 (b), is obtained using an Applied Biosystems automated DNA sequencer and consists of a two-dimensional array of intensity values. In this signal, the significant features are regions of high-intensity (shown shaded). In this paper, we confine our attention to the first sizing signal alone. We have described the automatic interpretation of the second sizing signal in (Pathak & Perlin 1994b).

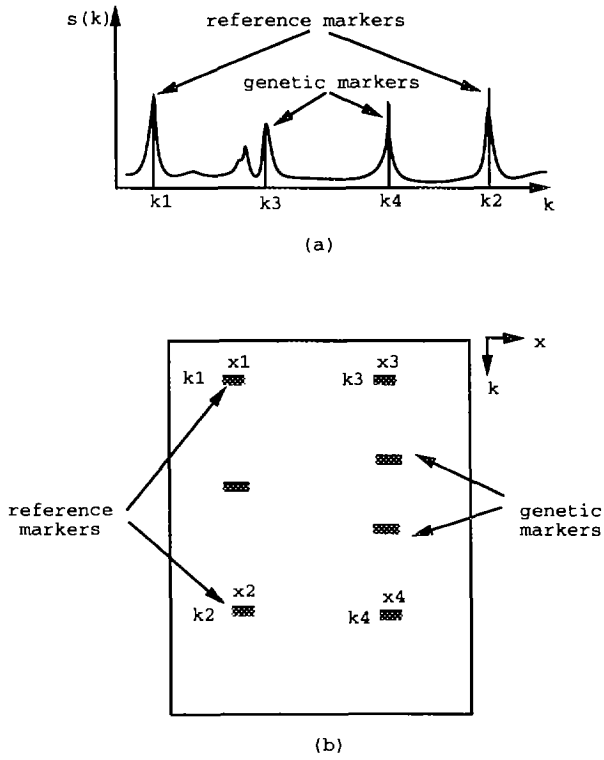

Figure 3: The two types of sizing signals.

The interpretation problem for sizing signals is the detection and labeling of the significant signal features. To solve the interpretation problem, a grid of reference marker variables is instantiated. It is illustrated in Figure 4. One grid dimension corresponds to a gel lane. The other grid dimension corresponds to refer-

R1: If the reference marker is r1 then find peaks between k1 and k2 of the sizing signal and compute the area underneath each peak.

R2: IF the reference marker is r1 AND possible peaks have been found for r1 THEN set the location of the reference marker variable for r1 to be the position of the peak with the largest area.

R3: IF the reference marker is r2 THEN find peaks between k3 and k4 of the sizing signal and compute the area underneath each peak.

R4: IF the reference marker is r2 AND the possible peaks have been found for r2 THEN set the location of the reference marker variable for r2 to be the position of the peak with the largest area.

Table 1: The rules to detect the reference markers in the sizing signal shown in Figure 3 (a).

ence markers in a gel lane. We assume that the same number of reference markers are present in each gel lane. Therefore, each grid variable corresponds to a reference marker in a specific gel lane. A grid variable is represented as an object with an associated set of attributes. The key attribute is the location of the reference marker given by the distance travelled by the reference marker. In Figure 3, these distances are given by k_i. A second attribute of a reference marker variable is its size in base pairs. This information is obtained from the user input. Thus, once the reference markers in a lane have been found, a characterization of migration distance on the gel as a function of the size of the product is available. To find the location of reference markers in the sizing signal, rule-based processing is used. Some of the rules for the sizing signal, shown in Figure 3 (a), are given in Table 1. Other rules (not shown here) are used when the reference marker peaks happen to lie outside the expected ranges on the gel.

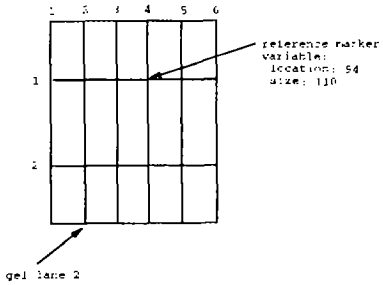

Figure 4: The grid representation.

After the reference markers have been located, the system finds the the possible alleles in the sizing signal for each lane. The markers whose PCR products are run in a given lane are known from the user input. Each gel lane is represented as an object with the attributes *markers* and *individual-name*. For each marker associated with a gel lane, a marker variable is instantiated and added to the marker attribute of the gel lane. Each marker variable is an object with the attribute *possible alleles*. Each possible allele is represented as an object with the attributes *location* and *area*. The object-based knowledge representation is illustrated in Figure 5.

R5: IF the locus is l1 THEN find peaks between k5 and k6 of the sizing signal and compute the area underneath each peak.

R6: IF the locus is l1 AND possible peaks have been found for l1 AND individual is male AND l1 is on X-chromosome THEN add a possible allele whose location is set to the peak with the largest area.

R7: IF the locus is l1 AND possible peaks have been found for l1 AND individual is female THEN add two possible alleles whose locations are set to the two peaks with the largest area.

Table 2: The rules to detect possible alleles in the sizing signal shown in Figure 3 (a).

To find the possible alleles for a given locus, rule-based processing is used. Some of the rules for the sizing signal, shown in Figure 3 (a), are given in Table 2.

Once the two enclosing reference markers for a given possible allele are obtained, a linear interpolation rule is used to determine the size of the possible allele in base pairs. The interpolation formula is: $size(a) = size(r1) + \frac{size(r2)-size(r1)}{location(r2)-location(r1)}(location(a) - location(r1))$.

Constraint Satisfaction of Genetic Flow

The key concept for diagnosis is that individuals who share common chromosomal content in the B/DMD region carry the same B/DMD gene, with high probability. For example, if an individual A at unknown risk shares B/DMD region content with a known affected individual B, then A most probably has a disease B/DMD gene. Chromosomal content is determined by comparing genetic markers (with high PIC values) that sample chromosome locations throughout the region of interest. If for every marker in a region the alleles of individuals A and B are identical, then A and B are inferred to have common chromosomal content in that region.

Note the distinction between genotypes and haplotypes. A genotype provides only unordered allele information: for each heterozygotic marker, it is not immediately evident which allele lies on which of the two chromosomes. That is, the *phase* of the allele is not known. A haplotype, however, provides ordered, phase-known information which assigns each allele to

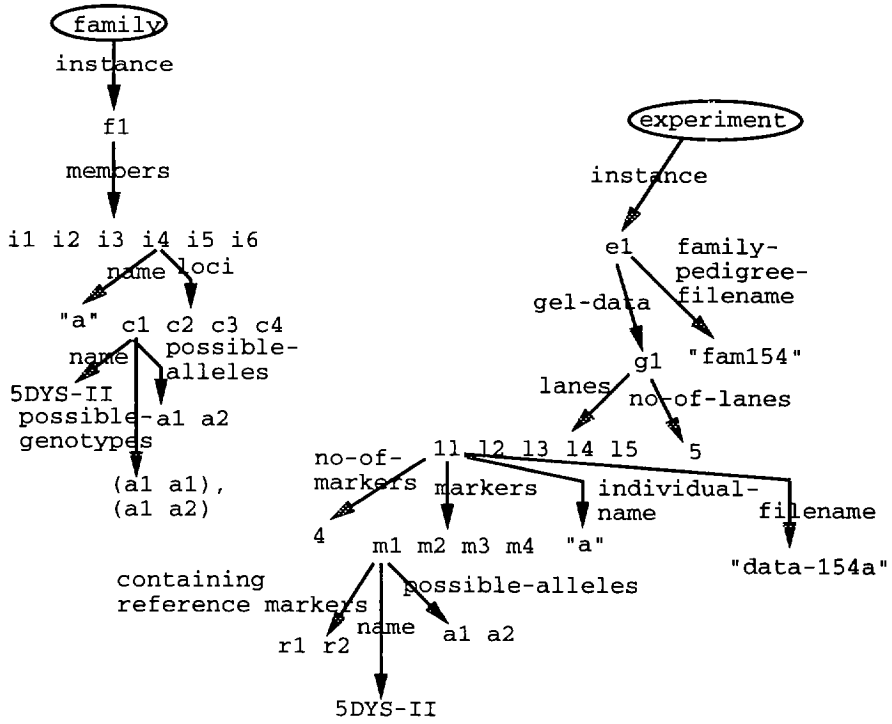

Figure 5: The knowledge representation for the sizing signal interpretation.

its proper chromosome.

The shared chromosomal content of the B/DMD gene within a family can be determined by tracing the genetic flow of haplotypes. Our system's second module computes this genetic flow by performing a constraint satisfaction for consistent haplotypes on the pedigree graph, thereby transforming genotype information into haplotype information. The constraint satisfaction begins at a male offspring, which is hemizygotic (having just one X chromosome), hence already haplotyped. The propagation explores the pedigree graph, locally visiting neighboring individuals one X-inheritance link away. When control passes from a neighbor's node to an individual's node, the local haplotyping computation at the node applies the following rules. Each marker genotype is analyzed separately:

- A male individual is assigned the haplotype of his hemizygotic genotype.
- When a female individual is set from a (haplotyped) male neighbor, the first haplotype is assigned the closest alleles matching the male's haplotype. The second haplotype is assigned the difference between the individual's genotype and the first haplotype.
- When a female individual is set from a haplotyped female neighbor, the first haplotype is assigned the closest alleles matching a haplotype of the neighbor. The second haplotype is assigned the difference between the individual's genotype and the first haplotype.

The graph traversal propagates only to unhaplotyped neighbors, so the algorithm terminates once consistent haplotyping is achieved.

Meiotic recombination events within the gene region can produce non-unique haplotype signatures. To check for recombination, independent graph propagations from different male descendants are done. The propagation locally terminates at an individual when a parent-child haplotype inconsistency is detected. This early termination suggests where recombination (or other mutation events) occur in the pedigree, and how to correct for their occurrence.

This constraint satisfaction approach is constructive, in that consistent haplotypes are developed, with the possibility of backtracking for error recovery. A different, destructive computational approach to haplotype constraint satisfaction is Waltz filtering (Waltz 1972). Here, the haplotypes for each individual are preenumerated, and the consistent arcs between them are listed. Filtering the arcs to preserve just the consistent haplotypes effects haplotyping. Although the number of haplotypes grows exponentially with the number of markers, with four markers the filtering approach is quite efficient.

Risk Assessment

One approach we use in our system is based on symbolic computation. The haplotypes can serve as signa-

tures of shared chromosomal content in the B/DMD gene region. Let H be the haplotype (i.e., the allele values of the four intragenic markers) of an affected individual. Then all males in the family with haplotype H are inferred to be affected, all females with haplotype H on one X chromosome are inferred to be carriers, and the remaining individuals are unaffected/non-carriers. In the presence of recombination, mutation, and gonadal mosaicism, a more refined Bayesian probabilistic approach can be used, as described next.

Once the genotypes for individuals in a family are obtained it is possible to compute carrier and affected-status probabilities for at-risk individuals in the family. A blackboard-style problem solver is used to automatically compute these risks (Pathak & Perlin 1994a). It is illustrated in Figure 6. The system interprets the family information and the genotype and phenotype data to generate all possible explanations for the observed data. Each explanation is an assertion about the phase at each individual and the presence or absence of disease mutation. The posterior probability for each explanation is calculated by computing its a priori and conditional probabilities. By design, the processing in the system is closely patterned after the processing used by skilled genetic counsellors.

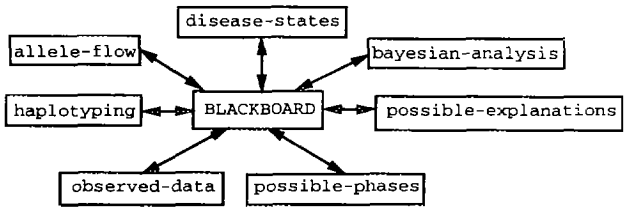

Figure 6: The risk-assessment module of the molecular diagnostics system.

There are seven key knowledge sources making up the system. Each knowledge source consists of a set of rules. The scheduling of knowledge sources is directed by a set of control rules. The *allele-flow* knowledge source reasons about the flow of alleles from parents to their children. The *disease-states* knowledge source reasons with the clinical data to infer whether or not an individual has the disease phenotype. The *haplotyping* knowledge source reasons with the flow of sequences of alleles from parents to their children. The *observed-data* knowledge source determines what data constitutes the observables. The *possible-phases* knowledge source generates the different phases given an individual's genotypes and subject to constraints obtained from the allele-flow and haplotyping analyzes. The *possible-explanations* knowledge source generates explanations using the phase assignments produced by possible-phases. Finally, the *Bayesian-analysis* knowledge computes the posterior probabilities of each explanation using disease knowledge.

Using the System for Automated Diagnosis

In this section we describe a complete example of the use of our intelligent molecular diagnostics system for genetic counseling a family with DMD. The family pedigree is shown in Figure 8. Family members are shown labeled with alphabets as well as in the more traditional genetic counseling notation using Roman numerals. The grandmother, D, is known to be heterozygote for DMD. Multiplex PCR amplification of four STR loci within the DMD gene is done for each assayed individual, and the sizing signals are obtained.

Phase 1A. The sizing signals for each individual is interpreted by the rule-based system described in Section 3. To illustrate the processing, consider the sizing signal for individual A shown in Figure 7. The single most significant signal peaks in the reference marker regions are identified as the two reference markers r1 and r2. Then, the significant peaks in the regions corresponding to each marker loci are identified. Each signal peak is given a size label using the reference markers for calibration.

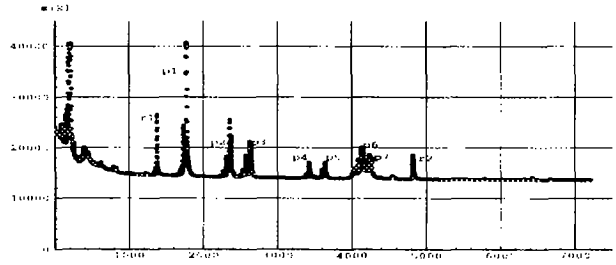

Figure 7: The sizing signal for individual A.

Phase 1B. The alleles for each marker of the male individuals B and E are determined by recording the DNA size corresponding to the largest peak in the sizing signals. This genotyping is shown in Figure 8.

For the female individuals A, C, and D, there are three cases: (1) Homozygotic alleles, as in A's (131 131), again record the DNA size of the largest peak in the sizing signals. (2) Heterozygotic alleles that have very different sizes, as in A's (158 170), are processed as two separate hemizygotic signals. (3) Heterozygotic alleles that have very similar sizes, as in D's (235 239), are determined by a consistency rule: superimposing

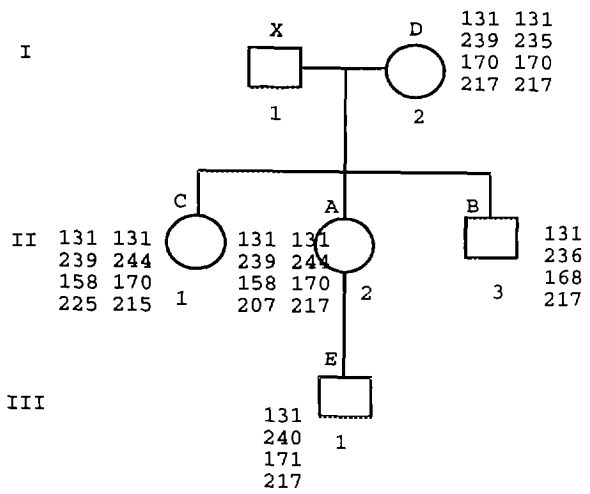

Figure 8: Allele Determination.

the correct two allele's sizing signals must generate the observed data. This is implemented by mathematically deconvolving D's complex sizing signal with a male's (e.g., E) hemizygotic sizing signal.

Phase 2. The haplotypes of each individual are determined by the constraint propagation described in Section . The X chromosome inheritance paths implicit in the pedigree form the arcs in the graph, as shown in Figure 9.

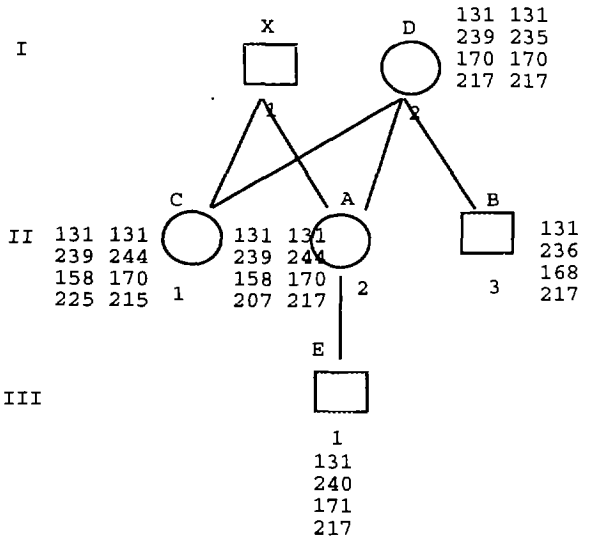

Figure 9: Haplotype Determination.

Propagating from the (haplotype known) male E, the nodes are visited in the order E, A, D, B, and C. At each step, the haplotyping from the neighbor is done. For example, consider female A's genotype G(A) = ((131 131) (239 244) (158 170) (207 217)). One of A's haplotypes is determined by identifying the alleles H1(A) = (131 239 171 217) which closely matches E's haplotype H(E) = (131 240 171 217). The alleles that remain, H2(A) = (131 244 158 207), comprise the second haplotype of A.

Phase 3. The clinical phenotype status of carrier or affected is now determined and the result is depicted in Figure 10.

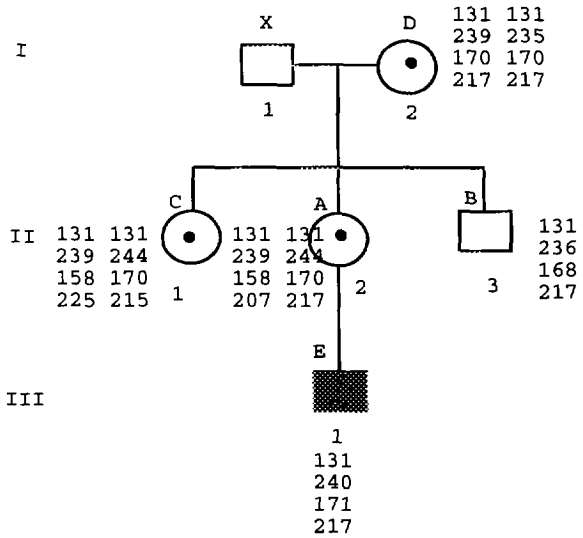

Figure 10: Phenotype determination.

In this figure, an affected individual is shown shaded, an unaffected individual is shown without shading, and a female carrier of the disease is shown with a black dot. Since D is known to be a carrier, and B is known to be unaffected, the haplotype that D does not share with B is presumed to be the DMD-containing disease haplotype, i.e., H1(D) = (131 239 170 217). This matches haplotypes in female individuals A and C, who are therefore inferred to be carriers, and in male individual E, who is thus inferred to be affected with DMD. The symbolic computation is done by approximate matching of the haplotype signatures. A probabilistic approach that provides confidence measures, as described next, can prove even more useful to the genetic counselor. Note that the pedigree figures are taken from our interactive user interface, and were automatically generated by the system for explanatory purposes.

The risk assessment module described in Section 5, exhaustively generates all possible explanations consistent with the observed data. For the case under consideration the explanation that has the mother of E to be a carrier is shown in Table 3. The posterior probability of this explanation is 0.99. Since, E has the disease haplotype, E's probability for being affected is also 0.99. The probabilistic risk assessment provides full accounting of possible recombinations and new mutations.

(and (and (carrier D) (phase ((131 239 170 217) (131 235 170 217)))) (and (carrier A) (phase ((131 239 170 217) (131 244 158 207)))) (and (carrier C) (phase ((131 239 170 215) (131 244 158 225)))) (disease-haplotype (131 239 171 217)))

Table 3: The explanation with the most significant probabilistic weight.

Conclusion

We have developed a intelligent prototype system for molecular diagnostics that automates the following knowledge-intensive tasks: (1) Rule-based interpretation of sizing signals; (2) Constraint satisfaction to determine genetic information flow; (3) Blackboard Bayesian problem solving to compute disease risk. Taken together, these modules provide a first (and crucial) step towards removing the analysis and interpretation bottlenecks that prevent efficient and accurate molecular diagnostics.

Our system draws on a large repertoire of AI techniques, including rule-based systems, machine vision, constraint satisfaction, blackboard problem solving, and Bayesian inference. Further, the molecular diagnostics problems mandate the use of such techniques to adequately represent and replicate the underlying human expertise. Thus, our system underscores the necessity of using AI methods to solve difficult real-world problems.

While the current system is still in prototype form, modules are finding use in the laboratories of our molecular biology collaborators. For example, the allele determination module is being used in the design of new genotyping experiments. Similarly, our module for knowledge- based interpretation of gel-based images is currently being extended to work with other detection systems in the molecular biology laboratory, such as gridded filter hybridizations. Once deployed, our system is expected to at least halve the time spent in the routine molecular diagnosis of DMD.

When fully operational, our system will enable much additional molecular biology research. Automated intelligent analysis and interpretation, coupled with high-throughput data generation experiments, will provide a feasible approach to rapid, accurate, and cost-effective disease gene localization, genome map construction, exploration of the biology of recombination, and molecular diagnostics of both sex-linked and autosomal diseases. Such intelligent system architectures provide molecular biologists with useful enabling tools to extend the capabilities of laboratory research.

References

Applied Biosystems, Foster City, California. 1993. *GS Analysis 1.2.0 Sequencer Analysis Program.*

Bibel, W. 1988. Constraint satisfaction from a deductive viewpoint. *Artificial Intelligence* 36:401–413.

Botstein, D.; White, R. L.; Skolnick, M.; and Davis, R. W. 1980. Construction of genetic linkage map in man using restriction fragment length polymorphism. *American Journal of Human Genetics* 32:314–331.

Chamberlain, J. S.; Gibbs, R. A.; Ranier, J. E.; and Caskey, C. T. 1990. Multiplex pcr for the diagnosis of duchenne muscular dystrophy. In Innis, M.; Gelfand, D.; Sninsky, J.; and White, T., eds., *PCR protocols: a guide to methods and application.* Academic Press. 271–281.

Emery, A. E. 1988. *Duchenne Muscular Dystrophy.* Oxford University Press.

Engelmore, R., and Morgan, T. 1988. *Blackboard Systems.* New York: Addison Wesley.

Hayes-Roth, F. 1985. Rule-based systems. *Communications of ACM* 28:921–932.

Koenig, M.; Hoffman, E. P.; Bertelson, C. J.; Monaco, A. P.; Feener, C.; and Kunkel, L. M. 1987. Complete cloning of the duchenne muscular dystrophy (dmd) cdna and preliminary genomic organization of the dmd gene in normal and affected individuals. *Cell* 50:509–517.

Lesser, V. R.; Fennell, R. D.; Erman, L. D.; and Reddy, D. R. 1975. Organization of the hearsay-ii speech understanding system. *IEEE Transactions on Acoustics, Speech and Signal Processing* 23:11–23.

Pathak, D. K., and Perlin, M. W. 1994a. Automatic computation of genetic risk. In *Proceedings of the Tenth Conference on Artificial Intelligence for Applications, San Antonio, Texas*, 164–170. IEEE Computer Society Press.

Pathak, D. K., and Perlin, M. W. 1994b. Intelligent interpretation of pcr products in 1d gels for automatic molecular diagnostics. In *Proceedings of the Seventh IEEE Symposium on Computer-Based Medical Systems, Winston-Salem, North Carolina.*

Saiki, R. K.; Gelfand, D. H.; Stoffel, S.; Scharf, S. J.; Higuchi, R.; Horn, B. T.; Mullis, K. B.; and Erlich, H. A. 1988. Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. *Science* 239:487–491.

Schwartz, L. S.; Tarleton, J.; Popovich, B.; Seltzer, W. K.; and Hoffman, E. P. 1992. Fluorescent multiplex linkage analysis and carrier detection for duchenne muscular dystrophy. *American Journal of Human Genetics* 51:721–729.

Waltz, D. L. 1972. Generating semantic descriptions from drawings of scenes with shadows. In Winston, P. H., ed., *The Psychology of Computer Vision.* McGraw-Hill.

Ward, P. A.; Hejtmancik, J. F.; Wirkowski, J. A.; Baumbach, L. L.; Gunnell, S.; Speer, J.; Hawley, P.; Tantravahi, U.; Caskey, C. T.; and Latt, S. 1989. Prenatal diagnosis of duchenne muscular dystrophy: Prospective linkage analaysis and retrospective dys-

trophin cdna analysis. *American Journal of Human Genetics* 44:270–281.

Watson, J. D. 1990. The human genome project: Past, present, and future. *Science* 248(4951):44–49.

Weber, J. L., and May, P. E. 1989. Abundant class of human dna polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* 44.