AI Magazine Volume 20 Number 1 (1999) (© AAAI)

Knowledge-Based Avoidance of Drug-Resistant HIV Mutants

Richard H. Lathrop, Nicholas R. Steffen, Miriam P. Raphael, Sophia Deeds-Rubin, Michael J. Pazzani, Paul J. Cimoch, Darryl M. See, and Jeremiah G. Tilles

■ We describe an AI system (CTSHIV) that connects the scientific AIDS literature describing specific human immunodeficiency virus (HIV) drug resistances directly to the customized treatment strategy of a specific HIV patient. Rules in the CTSHIV knowledge base encode knowledge about sequence mutations in the HIV genome that have been found to result in drug resistance to the HIV virus. Rules are applied to the actual HIV sequences of the virus strains infecting the specific patient undergoing clinical treatment to infer current drug resistance. A rule-directed search through mutation sequence space identifies nearby drug-resistant mutant strains that might arise. The possible combination drug-treatment regimens currently approved by the U.S. Food and Drug Administration are considered and ranked by their estimated ability to avoid identified current and nearby drug-resistant mutants. The highest-ranked treatments are recommended to the attending physician. The result is more precise treatment of individual HIV patients and a decreased tendency to select for drug-resistant genes in the global HIV gene pool. Initial results from a small human clinical trial are encouraging, and further clinical trials are planned. From an AI viewpoint, the case study demonstrates the extensibility of knowledge-based systems because it illustrates how existing encoded knowledge can be used to support new knowledge-based applications that were unanticipated when the original knowledge was encoded.

Human immunodeficiency virus (HIV) causes progressive deterioration of the immune system leading almost invariably to AIDS and death from opportunistic cancers and infections. Currently in the United States, it is estimated to infect 3 to 5 million persons, is the leading cause of death in adults from 14 to 35, and is the nation's leading cause of productive years of life lost aggregated over all age groups. HIV is estimated to infect 40 to 50 million persons worldwide (CDC 1997).

The high rate of HIV viral mutation both makes development of a vaccine difficult and results in rapid positive selection for drug-resistant mutant strains. Recent multidrug combination therapies are encouraging but in most cases ultimately fail because of the development of drug resistance (O'Brian et al. 1996). A general theory of HIV drug resistance still is not in hand, but a number of specific sequence mutations in the HIV genome have been described in the scientific literature and associated with increased resistance to certain drugs.

In this article, we describe an AI system (CTSHIV) intended to improve the clinical treatment of individual HIV patients by identifying drug resistance in advance and avoiding it in treatment. The improvement is accomplished by first identifying drug-resistant HIV mutant strains that already exist in the patient, or can be selected positively for, by certain treatments and then recommending a customized treatment strategy designed to avoid selection of such mutants. The result is more precise treatment of individual HIV patients and a decreased tendency to select for drug-resistant genes in the global HIV gene pool.

Project Goals

The project goals are to (1) connect knowledge contained in the scientific literature about HIV drug resistance directly to the treatment of individual HIV patients, (2) enable customized treatment strategies to be based on the HIV genotype that currently infects an individual

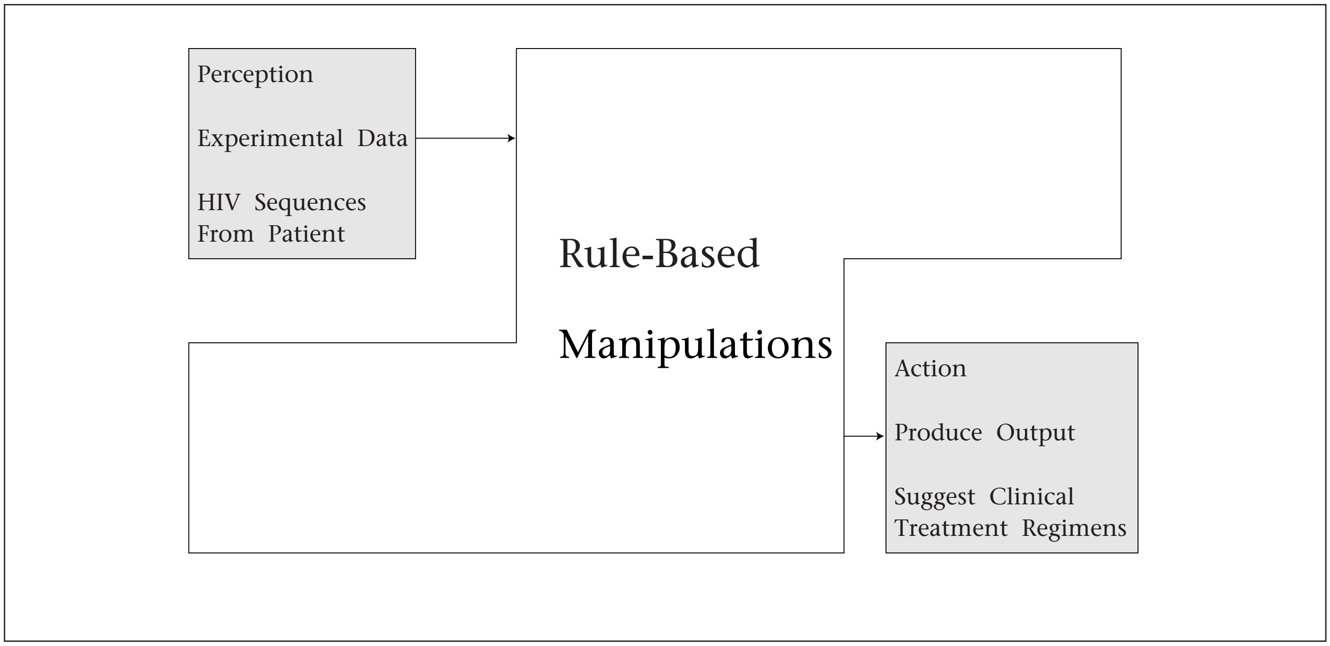


Figure 1. Rule-Based Manipulations by an Expert System That Connects Patient Data to Recommended Treatments Are a Central Foundational Pillar of AI.

HIV patient, (3) identify the nature and extent of drug resistance currently present in an individual HIV patient, (4) identify nearby drug-resistant mutant strains that could be selected positively for by some treatments, (5) rank the possible U.S. Food and Drug Administration (FDA)–approved treatments by an estimate of their ability to avoid both current and nearby drug-resistant mutants, (6) estimate the costs of the highest-ranked treatments, and (7) recommend treatments that are heuristically estimated to avoid known HIV drug resistance.

Related Work

This work rests on a central foundational pillar of AI: rule-based expert systems instantiated in the medical domain (for example, Buchanann and Shortliffe [1984]) (figure 1). For many such systems, a common diagnosis task is to identify the organism, from which treatment follows straightforwardly. Here, the organism is known to be HIV, but the treatment task is complicated by selective drug resistance.

Several AI applications have targeted HIV. An expert system based on experimental data from HIV patients (immunologic markers) has been used to diagnose the opportunistic non–Hodgkin's lymphomas that often develop (Diamond et al. 1994). Knowledge-based systems have been applied to HIV patient medical record systems (Safran et al. 1996; Musen et al. 1995), the monitoring of ongoing HIV patient protocols (Musen et al. 1996; Tu et al. 1995; Sobesky et al. 1994; Sonnenberg, Hagerty, and Kulikowski 1994), and HIV patient assessment (Xu 1996; Ohno-Machado et al. 1993). Less closely related are knowledge-based systems that apply qualitative modeling and process simulation to HIV laboratory systems (Ruggiero et al. 1994; Sieburg 1994). To our knowledge, CTSHIV is the first system to use HIV sequence data from HIV patients to estimate current and nearby drug-resistant mutants and recommend treatment combinations to avoid both.

Problem Description

The information content of an HIV virus is contained in a set of genes encoded in its genome. Each gene is a sequence of bases or nucleotides of four varieties. A gene can be represented as a string over an alphabet of four characters, one character representing each nucleotide. The HIV genome ultimately causes the production of gene products, often proteins, important in the virus life cycle. A *protein* is a sequence of amino acid residues of 20 varieties and can be represented as a string over an alphabet of 20 characters. Each amino acid in the protein is encoded by a block of three adjacent nucleotides in the gene, called a *codon*. Thus, the gene's nucleotide sequence produces the protein's amino acid sequence, which folds into a three-dimensional protein structure. Different amino acid types have dif-

```
CCC/ATT/AGC/CCT/ATT/GAG/ACT/GTA/CCA/GTA/AAA/TTA/AAG/CCA/GGA/ATG/GAT/GGC/CCA/AAA/GTT/AAA/CAA/TGG/CCA/  25
TTG/ACA/GAA/GAA/AAA/ATA/AAA/GCA/TTA/GTA/GAA/ATT/TGT/ACA/GAG/ATG/GAA/AAG/GAA/GGG/*AA/ATT/TCA/AAA/ATT/  50
GGG/CCT/GAA/AAT/CCA/TAC/AAT/ACT/CCA/GTA/TTT/GCC/ATA/AAG/AAA/AAA/GAC/AGT/ACT/AAA/TGG/AGA/AAA/TTA/GTA/  75
GAT/TTC/AGA/GAA/CTT/AAT/AAG/AGA/ACT/CAA/GAC/TTC/TGG/GAA/GTT/CAA/TTA/GGA/ATA/CCA/CAT/CCC/GCA/GGG/TAA/ 100
AAA/AAG/AAA/AAA/TCA/GTA/ACA/GTA/CTG/GAT/GTG/GGT/GAT/GCA/TAT/TTT/TCA/GTT/CCC/TTA/GAT/GAA/GAC/TTC/AGG/ 125
AAG/TAT/ACT/GCA/TTT/ACC/ATA/CCT/AGT/ATA/AAC/AAT/GAG/ACA/CCA/GGG/ATT/AGA/TAT/CAG/TAC/AAT/GTG/CTT/CCA/ 150
[CAG]/GGA/TGG/AAA/GGA/TCA/CCA/GCA/ATA/TTC/CAA/AGT/AGC/ATG/ACA/AAA/ATC/TTA/GAG/CCT/TTT/AGA/AAA/CAA/AAT/ 175
CCA/GAC/ATA/GTT/ATC/TAT/CAA/TAC/ATG/GAT/GAT/TTG/TAT/GTA/GGA/TCT/GAC/TTA/GAA/ATA/GGG/GAG/CAT/AGA/ACA/ 200
AAA/ATA/GAG/GAG/CTG/AGA/CAA/CAT/CTG/TTG/AGG/TGG/GGA/CTT/ACC/ACA/CCA/GAC/AAA/AAA/CAT/CAG/AAA/GAA/CCT/ 225
CCA/TTC/CTT/TGG/ATG/GGT/TAT/GAA/CTC/CAT/CCT/GAT/AAA/TGG/ACA/GTA/CAG/CCT/ATA/GTG/CTG/CCA/GAA/AAA/GAC/ 250
AGC/TGG/ACT/GTC/AAT/GAC/ATA/CAG/AAG/TTA/GTG/GGG/AAA/TTG/AAT/TGG/GCA/AGT/CAG/ATT/TAC/CCA/GGG/ATT/AAA/ 275
GTA/AGG/CAA/TTA/TGT/AAA/CTC/CTT/AGA/GGA/ACC/AAA/GCA/CTA/ACA/GAA/GTA/ATA/CCA/CTA/ACA/GAA/GAA/GCA/GAG/ 300
CTA/GAA/CTG/GCA/GAA/AAC/AGA/GAG/ATT/CTA/TAA/GAA/CAA/GTA/CAT/GGA/GTG/TAT/TAT/GAC/CCA/TCA/AAA/GAC/TTA/ 325
ATA/GCA/GAA/ATA/CAG/AAG/CAG/GGG/CAA/GGC/CAA/TGG/ACA/TAT/CAA/ATT/TAT/CAA/GAG/CCA/TTT/AAA/AAT/CTG/AAA/ 350
ACA/GGA/AAA/TAT/GCA/AGA/ATG/AGG/GGT/GCC/CAC/ACT/AAT/GAT/GTA/AAA/CAA/ATA/ACA/GAG/GCA/GTG/CAA/AAA/ATA/ 375
ACC/ACA/GAA/AGC/ATA/GTA/ATA/TGG/TGA/AAG/ACT/CCT/AAA/TTT/AAA/CTG/CCC/ATA/CAA/AAG/GAA/ACA/TGG/GAA/ACA/ 400
TGG/TGG/ACA/GAG/TAT/TGG/CAA/GCC/ACC/TGG/ATT/CCT/GAG/TGG/GAG/TTT/GTT/AAT/ACC/CCT/CCC/ATA/GTG/AAA/TTA/ 425
TGG/TAC/CAG/TTA/GAG/AAA/GAA/CCC
```

Figure 2. The Genomic Sequence of HIV Reverse Transcriptase (RT) Extracted from HIV Patient AA.

Each letter (A, C, G, T) represents a nucleotide; * represents any nucleotide. Each group of three letters represents a codon, set apart by slashes and counted by the numbers at the end of each line. The value of codon number 151 (CAG, bracketed) is the first three letters of line 7. This sequence encodes a three-dimensional protein structure similar to that shown in figure 3a but differing from it structurally to some extent, as specified by mutations in the sequence.

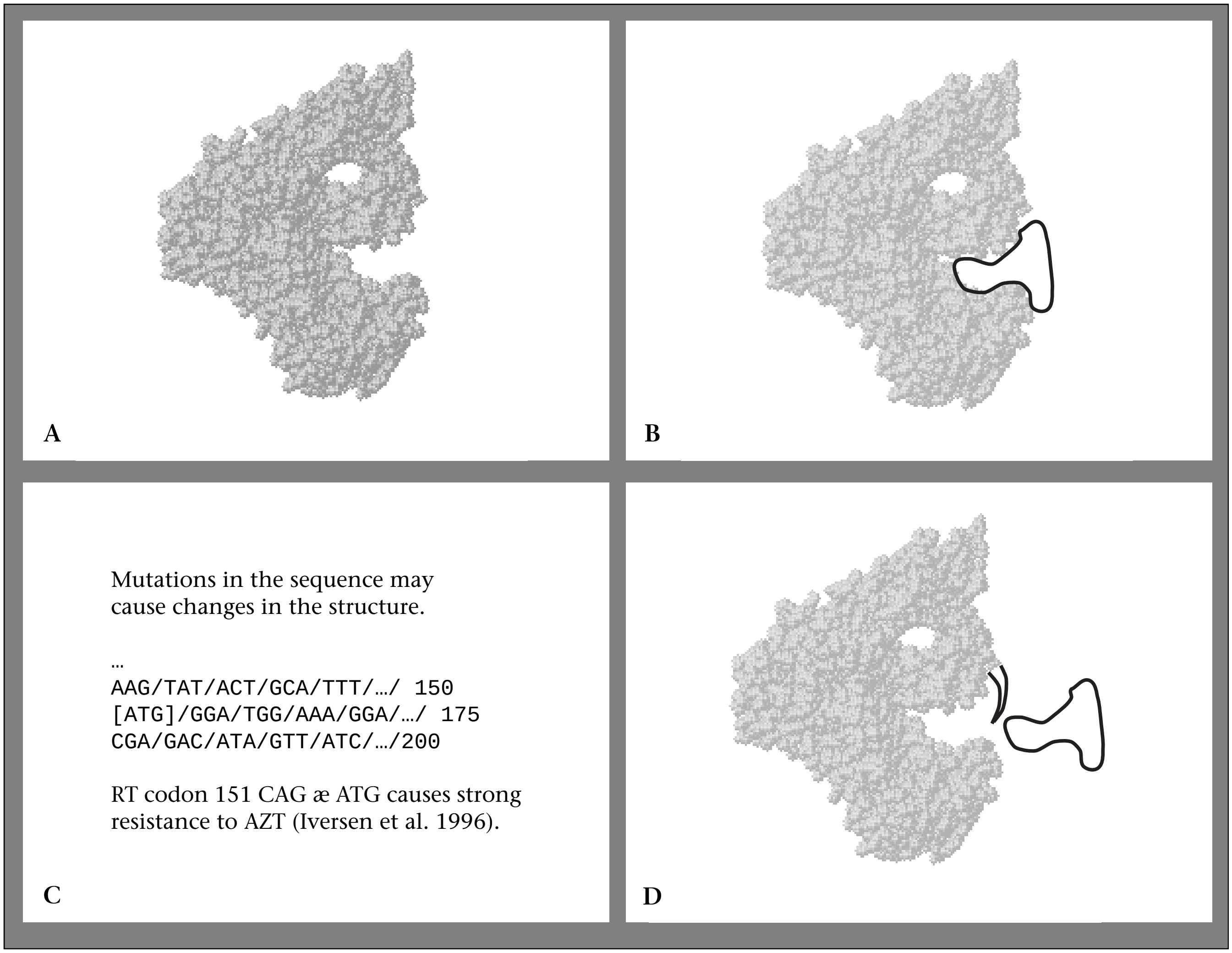
ferent sizes, shapes, and properties. Consequently, different gene sequences encode different protein structures.

The two proteins targeted by current FDA-approved drugs are called *reverse transcriptase* (RT) and *protease* (PRO). Figure 2 shows the string representation of an HIV gene for RT, and figure 3a shows a three-dimensional RT protein structure (Hsiou et al. 1996). The value of each codon in figure 2 controls the size, shape, and properties of a small local blob of structure in figure 3a.

The genome string must be copied from one generation to the next during the virus life cycle. Copying errors occur frequently and are called *mutations*. Mutations can change the structure of the virus and thus alter its function, or how it interacts with its environment. Mutant strains with genome sequences similar to the patient's current strain (close in Hamming or edit distance) appear spontaneously and continuously. In a full-blown case of AIDS, it is estimated that every single point mutation appears every day, every coordinated pair of point mutations appears once or more during the course of the infection, and even coordinated triples of point mutations can appear (Condra et al. 1995). The rapid mutation of the virus results in a population of related virus strains called a *quasispecies,* often consisting of a dominant strain and several minority strains.

A drug typically works by blocking a key part of the virus life cycle (figure 3b). A drug-resistant mutation occurs when a copying error in the viral genome (figure 3c) so alters the virus that it can perform the targeted step of its life cycle even in the presence of the drug (figure 3D). In the continued presence of the drug, the mutant strain can outcompete the dominant strain and thereby can itself become the dominant strain in the patient. This competition is often called *selective drug resistance* because the resistant mutant is selected for by the drug's presence. If unrecognized, the current treatment can lose its effect, and the patient's condition can deteriorate. The resulting strain is more challenging to treat because the treatment options have been reduced. If the drug treatment is changed in response, the potential is present for an additional drug-resistant mutation to develop. The use of an increasing variety of drugs has led to virus strains increasingly resistant to multiple drugs simultaneously. Sadly, the increasing prevalence of drug-resistant strains in the HIV global gene pool means that new patients can be infected by mutant strains that already have accrued resistance from previous hosts (Gu et al. 1994). Consequently, it is important to avoid selecting for drug- resistant mutants.

Combination treatments involving multiple drugs are one approach to avoiding drug resistance (Lange 1995). If the virus mutates to resist one drug but still is inhibited by another, it might be suppressed or unviable. In this case, the mutation cannot positively be selected for. Combinations can contain as many as four simultaneous drugs but usually do not



Mutations in the sequence may
cause changes in the structure.

```
…
AAG/TAT/ACT/GCA/TTT/…/ 150
[ATG]/GGA/TGG/AAA/GGA/…/ 175
CGA/GAC/ATA/GTT/ATC/…/200
```

RT codon 151 CAG æ ATG causes strong
resistance to AZT (Iversen et al. 1996).

Figure 3. The Function of the Virus Is Determined by Its Structure.

A. Molecular visualization of the three-dimensional structure of the HIV reverse transcriptase (RT) protein (PDB code 1DLO). Each sphere represents an atom. The structure is encoded in the HIV RT sequence (figure 2). B. A fanciful cartoon illustrating a drug binding to the structure and blocking a key part of the virus life cycle. The drug binds to the "mouth" of the "demon," so the "demon" cannot eat. C. Mutations in the sequence cause changes in the number, type, or spatial arrangement of atoms in the structure. D. A fanciful cartoon illustrating how a change in the structure can prevent a drug from binding. A "fang" appears, which blocks the drug's access to the mouth. Now the demon can eat undisturbed by the drug. Note that this resistance required changing only the size, shape, and properties of a single local blob of structure. This can be encoded by a single codon in the RT sequence and can be accessible mutationally. (Visualization by RASMOL.)

exceed three because of the potential for intolerable side-effects and toxicity. Severe side-effects often induce a patient to stop one or more drugs without knowledge of their physician, called *nonadherence* (formerly noncompliance). Nonadherence negates combination therapy and increases the likelihood of selecting for drug-resistant mutants.

Combinations containing at least one protease inhibitor are referred to as highly active antiretroviral therapy (HAART). HAART typically results in a dramatic drop in viral load within two weeks, often sustained for long periods of time. Enthusiasm for the potential of HAART to eradicate HIV has been tempered by the inevitable failure of these regimens because of the eventual development of drug resistance (Carpenter et al. 1996). The virus appears to remain in a proviral state in resting-memory T-cells, where it is inaccessible to antiretroviral drugs (Finzi et al. 1997; Wong et al. 1997). Mutations still can occur under HAART, although the mutation rate is greatly decreased (Jacobsen et al. 1996).

Nonetheless, knowledge of current or nearby mutants putatively resistant to one or more drugs is valuable to a physician treating an HIV patient. In conjunction with HAART, such

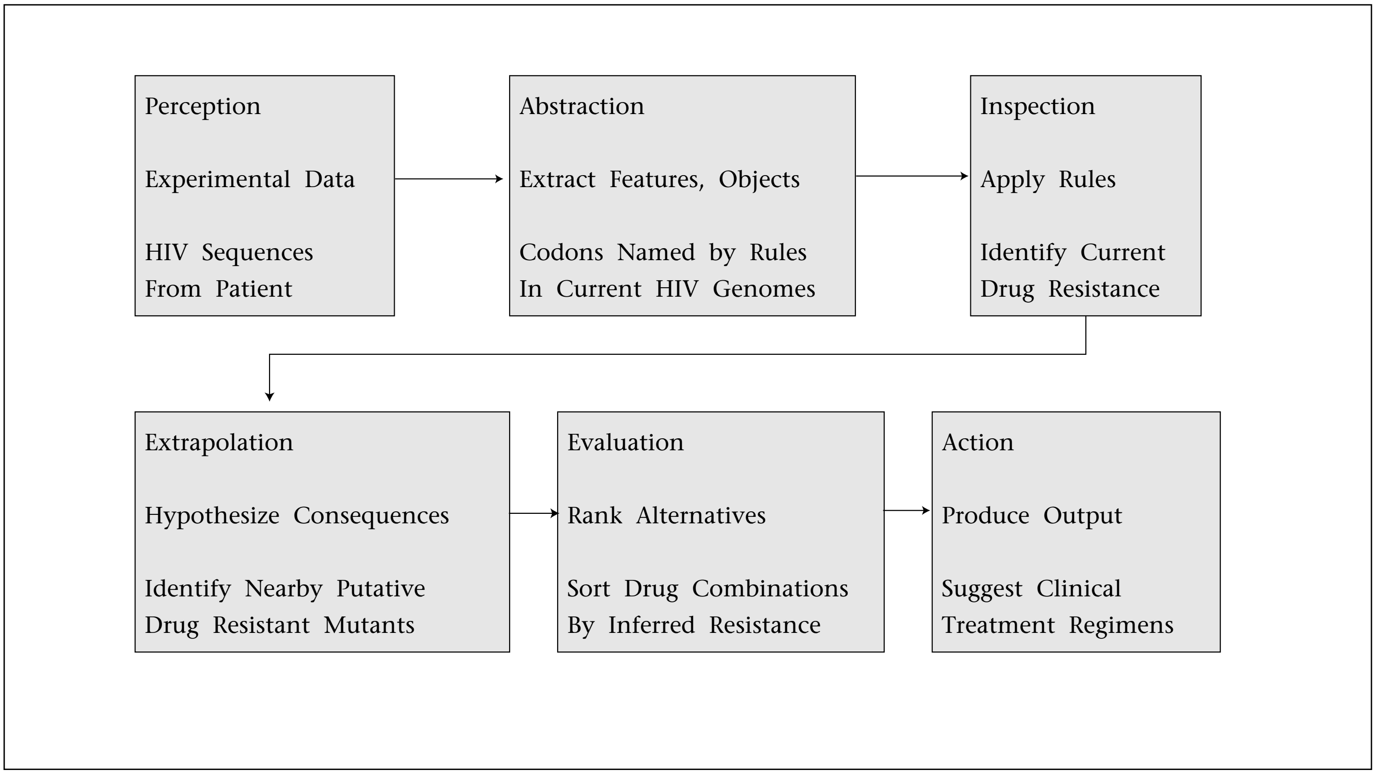


Figure 4. Application Overview Flowchart.

CTSHIV analyzes HIV sequences from the virus currently infecting a patient to suggest Federal Drug Administration–approved combination treatment regimens designed to avoid both current and mutationally close drug-resistant mutant strains. Processing the input HIV sequences involves identifying relevant sequence features, comparing them to the rule base to identify current resistance, exploring nearby mutation space to identify close drug-resistant mutants, ranking the possible treatments by their estimated ability to avoid resistance, and suggesting the highest-ranked treatment regimens to the attending physician.

knowledge can help select a combination of drugs less prone to be resisted. Currently, 11 drugs are approved by the FDA for HIV plus one available for compassionate use. These 12 drugs result in 407 different combination treatments of 4 or fewer drugs because some drugs should not be used together. A physician might find it tedious to scan many sequences, be unfamiliar with the latest HIV drug-resistant mutations reported, or have difficulty ranking the hundreds of treatment choices for each patient. CTSHIV mediates between the scientific literature and the patient's current infection to help a physician avoid HIV drug resistance.

Application Description

The application (1) accepts as input experimentally determined HIV sequences extracted from the patient, (2) extracts the relevant codons and constructs virtual genomes, (3) estimates current resistance by applying knowledge base rules, (4) searches nearby mutation sequence space to identify nearby putatively resistant mutants, (5) ranks the possible FDA-approved treatment regimens according to their ability to avoid selective drug resistance, and (6) recommends the highest-ranked treatment regimens to the attending physician. The application overview flowchart appears in figure 4. The application input-output constraints are shown in figure 5. A molecular visualization of the CTSHIV knowledge base appears in figure 6.

Patient's Experimental Data

The RT and PRO portions of the POL gene are amplified from each patient. Clones are produced, plasmid DNA is extracted, and the sequence is determined using a commercially available Applied Biosystems, Inc., sequencer. The RT sequence contains 1299 letters (433 codons), and the PRO sequence contains 297 letters (99 codons). Figure 2 shows an example HIV sequence from an HIV patient.

The sequences are prealigned to a standard reference HIV sequence, HXB2 (Fisher et al. 1985), using standard sequence-alignment algorithms. Deviations from the reference sequence correspond to mutations in the virus

Input from Patient

- 5 HIV clones (clone = RT + PRO)
- = 5 RT + 5 PRO (RT = 1299; PRO = 297)
- = 7980 letters of HIV genome

Possible FDA Treatments

- Possible drugs = 12 = 11 approved + 1 compassionate use
- $727 = \binom{12}{4} + \binom{12}{3} + \binom{12}{2} + \binom{12}{1}$ combinations
- –320 some drugs should not be used together
- = 407 possible FDA-approved combination treatments

FDA = U.S. Federal Drug Administration

Figure 5. Application Input-Output Constraints.

Although the problem is a reasonable size for a machine, a human would find the task tedious and error prone. Input consists of sequence data from five HIV clones taken from the patient. Each RT sequence contains 1299 letters and each PRO sequence contains 297 letters, for a total input of 7980 letters of HIV genomic information. Output consists of selecting recommended treatments from the Federal Drug Administration (FDA)–approved combination treatments available. Currently, there are 11 drugs approved by the FDA for HIV, plus 1 available for compassionate use. These 12 result in 407 different combination treatments of 4 or fewer drugs because some drugs should not be used together.

infecting the patient. Typically, 5 RT sequences and 5 PRO sequences, a total of 7980 letters of HIV genomic information, are the input experimental data on the patient's current infection.

Extract Features, Objects

Processing in this step is routine. The features extracted are exactly those codons in positions referred to by the antecedent of some rule. Other codon positions are not yet associated with known drug resistance. Currently, 55 rules mention 31 different codon positions, 20 in RT and 11 in PRO. HIV sequences are replaced by abstract objects consisting of only these codon positions. All possible virtual genomes are formed consistent with the experimental sequences.

Identify Current Resistance

Current drug resistance is identified by applying the 55 rules in the knowledge base to the HIV sequences from the patient. The rules represent knowledge about HIV drug resistance as a set of if-then rules of the form

IF < antecedent > THEN < consequent >
WITH < weight > (references).

For example, one such rule in CTSHIV is

IF the value of RT codon number 151 is ATG,
THEN infer resistance to AZT, ddI, d4T, and ddC
WITH weight = 1.0 (Iversen et al. 1996)

The weight associated with a rule is not a confidence, as in many expert systems. The rules are assumed to have a high degree of confidence because of the peer-review process and general integrity of the scientific literature. Rather, the weight reflects the estimated level of resistance to a particular drug and is part of the consequent. Different virus strains can resist a particular drug to different degrees, which is represented by weights that range from 0.1 (low resistance) to 1.0 (high resistance) based on expert advice and the level of resistance reported in the literature.

To estimate current resistance, rule weight is multiplied by the fraction of viral sequences that trigger the rule and combined additively. As a summary metric, we use

$$CurrWt(D) = \sum_{r \in Rules(D)} \sum_{s \in S} Apply(r,s)/\mid S \mid$$

where D is a set of drugs that make up a combination therapy, $Rules(D)$ are the rules that confer resistance to a drug in D, S is the set of the HIV sequences extracted from the patient, and $Apply(r, s)$ yields the rule weight of r if r fires on s and 0 if not. *CurrWt* is comparable only between combinations with the same number of drugs because any superset of a drug combination has equal or greater current weight. Under this model, the total current level of resistance to a multidrug combination is the sum of the current resistances to each drug. The effect of this computation is to identify drug combinations that have little or no current resistance and therefore attack the virus strongly.

Predict Nearby Resistant Mutants

Nearby resistant mutants are predicted by a backward-chaining search through mutation sequence space, beginning with the patient's current HIV sequences. At each step, a sequence that does not fire a rule is used to generate several new sequences that do. The new sequences are identical except that codon positions mentioned by the rule are modified so that the rule does fire. They represent mutants that are close in Hamming distance but resist the drugs mentioned by the rule. For example, figure 2 generates figure 3c this way. Conceptually, every virtual mutant within a predetermined Hamming distance cutoff is examined. Currently, all mutants up to and including Hamming distance three are considered.

To predict nearby mutants, rule weights are combined by taking the maximum across all mutants of the minimum resistance across all drugs in the combination. As a summary metric, we use

$$m\_dist(D) = min\{h \mid \exists x \in M(S, h), \forall d \in D, 0 < CurrWt(d, x)\}$$

$$m\_wt(D) = \max_{x \in M(s, m\_dist(D))} \min_{d \in D} CurrWt(d,x)$$

$$MutScore(D) = \max\{0, h_{\max} - m\_dist(D) + m\_wt(D)\}$$

where h_{max} bounds the maximum Hamming distance considered, $CurrWt(d, x)$ applies *CurrWt* to d using x instead of S, and $M(S, h)$ is the set of mutants of S at Hamming distance h. $m\_dist(D)$ is the minimum Hamming distance at which a mutant occurs that resists every drug in D, and $m\_wt(D)$ is the rule weight of the least resisted drug in D by the most resistant such mutant. *MutScore* is comparable between drug combinations with different numbers of drugs. $MutScore(D)$ is zero if no mutant within Hamming distance h_{max} of S resists every drug in D. Otherwise, its integer part is h_{max} minus the Hamming distance to such a mutant, and its fractional part is the maximum-minimum rule weight of such mutants.

Under this model, a drug combination suppresses a population of mutants only as strongly as it suppresses the most resistant mutant, and a mutant resists a drug combination only as strongly as it resists the least

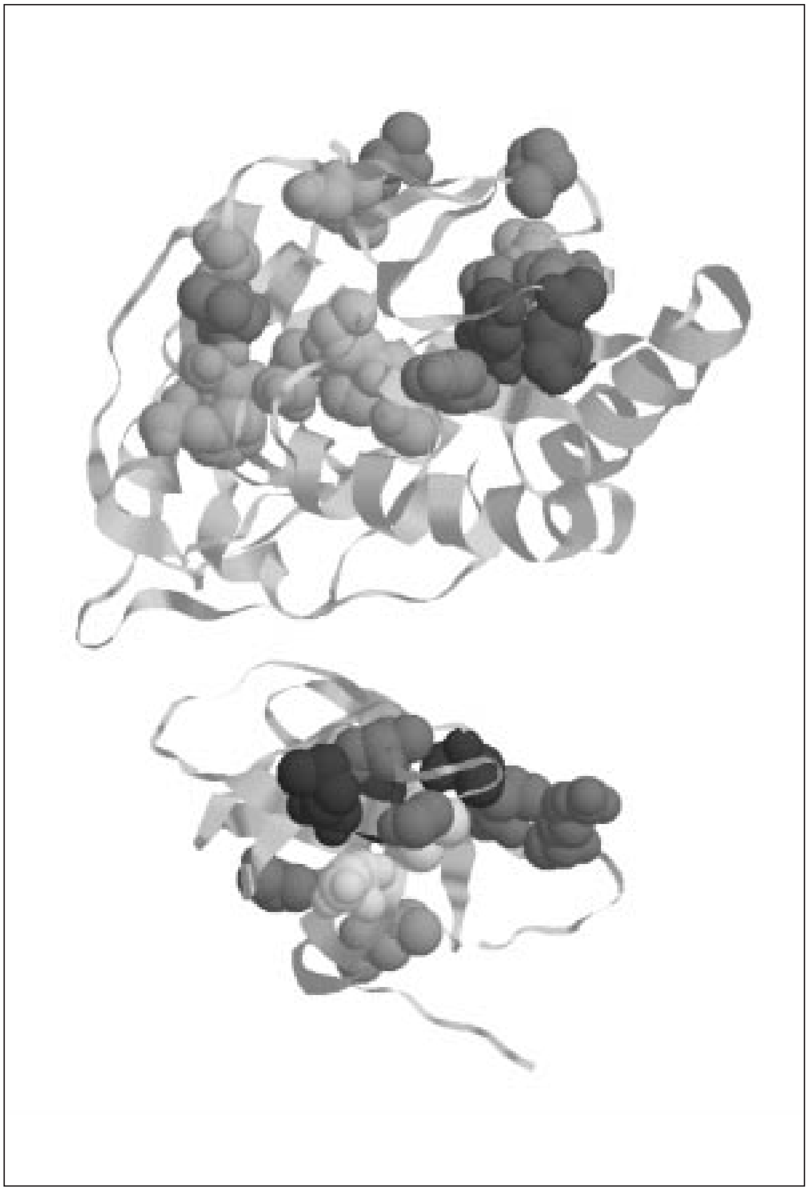


Figure 6. A Molecular Visualization of the CTSHIV Knowledge Base.

HIV protease and reverse transcriptase (palm and fingers) monomer backbones are green ribbons. Full-atom residues appear at locations mentioned by a rule antecedent.

The key enabling AI technology is knowledge representation of the relevant scientific literature about HIV drug resistance as a set of sequence-pattern rules on the HIV genome.

resisted drug in the combination. The effect of this process is to identify nearby mutants that resist every drug in a combination and drug combinations such that no nearby mutant resists every drug.

Rank Alternatives

CTSHIV ranks alternative drug combinations using the current resistance weight (*CurrWt*) and the nearby mutant resistances (*MutScore*). This ranking is done using any monotonic function *f* of *CurrWt* and *MutScore*. Currently, we use Euclidean distance

$$f(D) = \sqrt{CurrWt^2(D) + MutScore^2(D)}$$

to rank drug combination *D*. Values near or at zero indicate little or no resistance, and increasing positive values indicate increasing resistance. The best-ranked combinations represent a satisficing compromise along both metrics simultaneously.

Sketch of Ranking Algorithm The previous model gives rise to three nested optimization problems: (1) identify the drug combinations that most strongly suppress a population of mutants centered on the patient's current HIV strains; (2) for a given drug combination, identify the most resistant mutant in the population; and (3) for a given drug combination and mutant, identify the least resisted drug.

Because *CurrWt* is independent of nearby mutants, choosing any monotonic function for *f* guarantees for fixed *D* that the mutant strains that minimize *f* also minimize *MutScore*. Thus, if *x* is the most resistant mutant to *D* found so far, then the function $h(D) = f(D, x) = (CurrWt^2\ (D) + MutScore^2\ (D, x))^{1/2}$ is an admissible heuristic for *f*(*D*), where *f*(*D, x*) applies *f* to *D* using *x* instead of *S*.

Initially, objects are created to represent all FDA-approved drug combinations (currently 407). The one-, two-, three-, and four-drug combinations are treated separately. At each step, the current best *i*-drug combination is examined. Mutation sequence space is searched for a more resistant mutant than the most resistant mutant found so far. If a more resistant mutant is found, then the more resistant mutant replaces the previous most resistant mutant and the process iterates; otherwise, the drug combination is returned. This process repeats until enough highly ranked *i*-drug combinations are found (currently the best five plus the best RT-only combination). Indexing and branch-and-bound techniques avoid wasteful recomputation and prune unnecessary evaluations. Currently, CTSHIV runs in about a minute for each patient, which is acceptable for now.

Suggest Clinical Treatment Protocols The final result of application processing is to recommend the five highest-ranked combinations of one, two, three, and four drugs. The next-highest–ranked RT-only combination is shown for comparison. Figure 7 shows three-drug combinations recommended for an HIV patient. Figure 8 shows an example nearby resistant mutant. It is hoped that the CTSHIV output will increase patient adherence by clearly showing the deleterious effects of failing to take all medications. Figure 9 shows the projected consequences of nonadherence to the highest-ranked three-drug combination in figure 7. Other output produced, not shown here, includes the estimated current resistance to each drug individually, a list of cost codes and drug abbreviations, the rules that fired on the patient's current sequences and the associated citations in the scientific literature, and a detailed listing of differences between the patient sequences and a standard reference sequence.

Limitations

There are important limitations to the previous approach. Sequence-based rules capture only part of the domain knowledge about drug resistance, albeit a clinically useful part. Drug resistance can arise for other domain-specific reasons that cannot be represented easily as rules. More complicated organisms, such as bacteria and fungi, have more sophisticated resistance mechanisms than addressed here. Current sequencing techniques can provide only partial or no information about minority strains. The rule set is only as complete as current scientific knowledge allows. Currently, it might be possible to infer when resistance might occur based on genome sequences actually seen in the patient that correspond to resistance-conferring mutations described in the scientific literature. However, it is impossible to guarantee the nonexistence of an unsuspected resistant mutant.

Uses of AI Technology

The key enabling AI technology is knowledge representation of the relevant scientific literature about HIV drug resistance as a set of sequence-pattern rules on the HIV genome. Rule-based expert systems declaratively represent knowledge of a specialized problem and facts about a specific case and, from these, draw inferences about the case. Here, the rules encode information on drug-resistant mutations of HIV, the facts are the sequences

```
These protocols with 3 drugs are recommended:    CurrWt MutScor   0 Mut   1 Mut   2 Mut   3 Mut
 A5 SAQUINAVIR NELFINAVIR D4T:                    0.06     0.1     0.0     0.0     0.0     0.1
 B3 SAQUINAVIR DELAVIRDINE D4T:                   0.00     0.2     0.0     0.0     0.0     0.2
 C3 SAQUINAVIR NEVIRAPINE D4T:                    0.00     0.4     0.0     0.0     0.0     0.4
 D4 SAQUINAVIR DELAVIRDINE AZT:                   0.00     0.6     0.0     0.0     0.0     0.6
 E4 SAQUINAVIR NEVIRAPINE AZT:                    0.00     0.6     0.0     0.0     0.0     0.6
RF3 DELAVIRDINE DDI AZT:                          0.08     1.2     0.0     0.0     0.2     0.9
```

Figure 7. Example Three-Drug Output from HIV Patient AA, Showing a Favorable Resistance Profile.

For the highest-ranked treatment, current resistance (CurrWt) and nearby mutation score (MutScor) are small, and only a weakly resistant mutant appears even out to Hamming distance three (3 Mut). The letters A to F identify treatments. Treatment F is the best RT-only treatment (indicated by the prefixed letter R). Digits after the letters indicate cost codes (0 = $0 to $200, . . . , 3 = $600 to $800, 4 = $800 to $1000, 5 = $1000 to $1200, . . . , for each month estimated average wholesale cost).

```
                                                 CurrWt MutScor   0 Mut   1 Mut   2 Mut   3 Mut
 A5 D4T NELFINAVIR SAQUINAVIR:                    0.06     0.1     0.0     0.0     0.0     0.1
    Current: (NELFINAVIR)    RT 151:CAG->ATG by R11 (D4T)    PRO 90:TTG->ATG by R28
(SAQUINAVIR)
```

Figure 8. Example Output for HIV Patient AA Showing One Example of the Closest Mutants Inferred to Most Resist Every Drug in the Top-Ranked Three-Drug Combination of Figure 7.

Three letters must change simultaneously. Currently, Nelfinavir is resisted; changing two letters at RT 151 resists D4T, and changing one at PRO 90 resists Saquinavir.

```
                                                 CurrWt MutScor   0 Mut   1 Mut   2 Mut   3 Mut
 A5 SAQUINAVIR NELFINAVIR D4T:                    0.06     0.1     0.0     0.0     0.0     0.1
        If stop NELFINAVIR:                                0.6     0.0     0.0     0.0     0.6
        If stop SAQUINAVIR:                                1.1     0.0     0.0     0.1     1.0
        If stop D4T:                                       2.1     0.0     0.1     0.6     1.1
```

Figure 9. Example Output for HIV Patient AA Showing the Projected Result from Stopping Any Single Drug in the Top-Ranked Three-Drug Combination of Figure 7 (Saquinavir, Nelfinavir, D4T).

Mutants are closer or worse or both. Stopping Nelfinavir is bad, stopping Saquinavir is worse, and stopping D4T is worst of all.

of HIV genome obtained from a specific individual, and the inference to be drawn is a set of drug combinations to be recommended for the patient.

Rule forward chaining from the patient's current HIV sequences yields currently resistant HIV mutants. Rule backward chaining through sequence space yields the nearby putatively resistant mutants. Together, they allow CTSHIV to avoid both sets of mutants. AI heuristic-search methods speed the search. The intelligent-agent paradigm also proved useful as an organizing principle. Except for the lowest level (domain specific), figure 4 could represent any intelligent agent connecting perception to action.

Application Use and Payoff

The first HIV patient data was run through the CTSHIV system in June 1996. In February 1997, the application began its first round of human clinical trials on 14 HIV patients at the University of California at Irvine (UCI) and at the Orange County Center for Special Immunology as a satellite site, under the auspices of the California Collaborative Treatment Group (CCTG). Informed consent was obtained using a form approved by UCI Institutional Review Board. All patients had detectable viral load at baseline (mean $\log_{10}$ load of 4.67 ± 2.16) and failure of at least one previous antiviral treatment regimen because of the emergence of

Responders
- 9 = complete; no detectable viral load at completion
- 1 = partial; viral load reduction ≈ 25x at completion

Nonresponders
- 2 = treatment failure at completion
- 2 = withdrawn (1 death, 1 disappeared)

Of 14 Enrollees
- 64% = 9/14 had no detectable viral load
- 71% = 10/14 were responders

Of 12 Completers
- 75% = 9/12 had no detectable viral load
- 83% = 10/12 were responders

Table 1. Summary of Small-Scale Human Clinical Trials: Outcome of 14 Patients after 1 Year of Treatment.

Detailed per-patient trial outcomes are reported in Cimoch et al. (1998).

drug resistance. These patients, already expected to be infected by drug-resistant strains of HIV, are considered among the most challenging to treat.

Results from these small-scale trials are encouraging (Cimoch et al. 1998). As shown in table 1, 12 patients completed 1 year of trials (2 patients withdrew prior to completion). After 1 year of treatment, 9 patients who had failed at least 1 prior treatment regimen had an undetectable viral load (9 complete responders, 64 percent of enrollees, 75 percent of completers), and 1 other patient had ≈ 25x viral load reduction (10 partial responders, 71 percent of enrollees, 83 percent of completers). This improvement compares to about 20 percent in everyday practice in the same patient population. Note that in typical clinical trials, the percentage of viral-load undetectable patients diminishes over time. We expect and are seeing improvement over time based on CTSHIV-suggested treatment regimens. Detailed per-patient trial outcomes are reported in Cimoch et al. (1998).

Currently, a total of 68 HIV patients have been run through the CTSHIV system. A new round of CTSHIV (phase 2) clinical trials are under way, enrolling patients from UCI and Stanford University. Twenty patients will be enrolled in this open-label trial. The purpose is to evaluate whether the CTSHIV system can assist in the management of patients who have failed multiple antiretroviral regimens using standard monitoring. Collaborations with several other groups involved in the treatment of HIV patients have begun and are expanding. An Affymetrix gene chip machine has been purchased, and sequencing throughput will increase dramatically when it comes online. Because of the early encouraging results of the clinical trials, widespread recognition of the drug-resistance problem, and the high rate of HIV infection in the general population, we expect use of the application to increase sharply in the near future.

Application Development and Deployment

Three domain experts (Darryl See, Douglas Richman, Edison Schroeder) began extracting rules from the scientific literature in September 1995. The first rule set was completed in May 1996.

The first rule-based system prototype was developed to identify current resistance already present in the patient's HIV infection (Lathrop and Pazzani 1999; Pazzani, See, et al. 1997). It was coded in FOCL-1-2-3 (Pazzani and Kibler 1992), a Lisp-based expert shell. It was begun in March 1996 and completed in June 1996. It was recoded in JAVA between April and June 1997 (Pazzani, Iyer, et al. 1997).

The ability to use the rules to search mutation sequence space for nearby drug-resistant mutants was unanticipated when the original knowledge was encoded and the first prototype developed, thus demonstrating the robustness and extensibility of knowledge-based systems. A Lisp-based mutation space search engine was begun in November 1996 and completed in May 1997. The two subsystems were integrated and recoded in Lisp between October and December 1997. The application is deployed primarily by the e-mail exchange of input clinical data and output recommended treatments. We have developed an automatic e-mail server as well as a World Wide Web–based graphic interface to the e-mail server. The server extracts patient data from the body of an e-mail message, automatically enqueues the application to process it, and e-mails the results back to the sender. User interface enhancements will follow.

Deployment has been smooth largely because the application end users to date have been enthusiastic domain experts who are currently treating HIV patients. For cases where a treatment regimen has failed because of the development of drug resistance, the application enables them to base their next choice of treatment regimen on scientific principles and experimental data. This knowledge-based treatment selection replaces the blind intuition and guesswork that formerly guided treatment switches after treatment failure. They are glad to see their patients improve, anxious to see the application succeed, and tolerant of the few glitches.

Maintenance

It is doubtful that the knowledge base will be complete until HIV is eradicated. Maintenance of CTSHIV is equivalent to adding new rules from the scientific AIDS literature. The rules are revised by three domain experts every three months by extracting new rules that have appeared in the literature in the interim. Relevant articles are retrieved by keyword-based literature search, old rules revised as needed, and new rules composed manually.

In the future, we anticipate that the challenge of extending the knowledge base will provide fruitful opportunities for intelligent applications. An intelligent information-retrieval system could monitor the literature, retrieve papers that mention HIV drug- resistant mutations, extract candidate rules, and automatically enqueue review by domain experts. Other AI approaches could suggest when to test a patient strain for possible resistance to a specific drug. Predicting when a putative mutant is unviable and coping with resistance that occurs outside the rule set are further challenges for intelligent systems. Machine-learning and data-mining techniques could learn new rules, infer trends, and recognize regularities in resistance patterns.

Summary

We described CTSHIV, an AI application that connects the scientific literature describing specific HIV drug resistances directly to the HIV virus strain infecting a specific HIV patient. The application identifies current drug-resistant mutant strains by rule forward chaining from the patient's current HIV sequences and nearby putatively resistant mutants by rule backward chaining through mutation sequence space. It ranks the current FDA-approved treatment regimens according to their estimated ability to avoid both sets of resistant mutants and recommends a customized treatment strategy for the individual patient involved. Thus, the significance of the application is (1) a method for addressing HIV drug resistance in the clinic, especially treatment switches after treatment failure, based on scientific principles and experimental data; (2) a decreased tendency to select for drug resistance in the global HIV gene pool; and (3) a possible model for the use of knowledge-based systems in other drug-resistant viruses.

This article also illustrated the robustness and extensibility of knowledge-based systems. It showed how knowledge originally encoded to perform one knowledge-based task easily can be redirected to perform another, even one not anticipated when the original knowledge was encoded. This result supports knowledge base efforts to encode knowledge in societally important areas.

Acknowledgments

Douglas Richman and Edison Schroeder helped extract the original knowledge base from the AIDS scientific literature and revise the knowledge base every three months. Doug Cable and Winnie Huang cosupervised the first CTSHIV clinical trials. Data entry and analysis were done by Richard Haubrich and Allen McCutchan at the University of California at San Diego. Carol Kemper is cosupervising the next round of clinical trials. Tom Gingeras contributed useful practical insights and advice. We gratefully acknowledge the technical assistance of Tonya Clark and Marikel Chatard in cloning, PCR, and RNA-DNA extraction. Funding has been provided by Roche Molecular Systems, the California University–wide AIDS Research Program through the CCTG and the National Science Foundation under grant IRI-9624739. CTSHIV is available from University/Industry Research and Technology, 380 University Tower, University of California, Irvine, CA 92717.

References

Buchanann, B., and Shortliffe, E., eds. 1984. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, Mass.: Addison-Wesley.

Carpenter, C.; Fischl, M.; Hammer, S.; Hirsch, M.; Jacobsen, D.; Katzenstein, D.; Montaner, J.; Richman, D.; Saag, M.; Schooley, R.; Thompson, M.; Vella, S.; Yeni, P.; and Voldering, P. 1996. Antiretroviral Therapy for HIV Infection in 1996. *Journal of the American Medical Association* 276:146–154.

CDC. 1997. HIV/AIDS Surveillance Report. 9(1) ed. Atlanta, Ga.: Centers for Disease Control and Prevention.

Cimoch, P.; See, D.; Pazzani, M.; Reiter, W.; Lathrop, R.; Fasone, W.; and Tilles, J. 1998. Application of a Genotypic-Driven Rule-Based Expert Artificial Intelligence Computer System in Treatment Experienced HIV-Infected Patients. Immunologic and Virologic Response. Paper presented at the Twelfth World AIDS Conference, 28 June–3 July, Geneva, Switzerland.

Condra, J.; Schlief, W.; Blahy, O.; Gabryelski, L.; Graham, D.; Quintero, J.; Rhodes, A.; Robbins, H.; Roth, E.; Shivaprakash, M.; Titus, D.; Yang, H.; Teppler, K.; Squires, P.; Deutsch, P.; and Emini, E. 1995. In Vivo Emergence of HIV-1 Variants Resistant to Multiple Protease Inhibitors. *Nature* 374:569–571.

Diamond, L.; Nguyen, D.; Jouault, H.; Imbert, M.; and Sultan, C. 1994. An Expert System for the Interpretation of Flow Cytometric Immunophenotyping Data. *Journal of Clinical Computing* 22(1): 50–58.

Finzi, D.; Hermankova, M.; Pierson, T.; Carruth, L.; Buck, C.; Chaisson, R.; Quinn, T.; Chadwick, K.; Margolick, J.; Brookmeyer, R.; Gallant, J.; Markowitz, M.; Ho, D.; Richman, D.; and Siliciano, R. 1997. Identification of a Reservoir for HIV-1 in Patients on Highly Active Antiretroviral Therapy. *Science* 278(5341): 1295–1300.

Fisher, A.; Collaiti, L.; Ratner, R.; Gallo, R.; and Wong-Staal, F. 1985. A Molecular Clone of HTLV III with Biologic Activity. *Nature* 316(6025): 262–265.

Gu, Z.; Gao, Q.; Fang, H.; Parniak, M.; Brenner, B.; and Wainberg, M. 1994. Identification of Novel Mutations That Confer Drug Resistance in the Human Immunodeficiency Virus Polymerase Gene. *Leukemia* 8S1:5166–5169.

Hsiou, Y.; Ding, J.; Das, K.; Clark, A.; Hughes, S.; and Arnold, E. 1996. Structure of Unliganded HIV-1 Reverse Transcriptase at 2.7 Angstroms Resolution. *Structure* 4(7): 853–860.

Iversen, A.; Shafer, R.; Wearly, K.; Winters, M.; Mullins, J.; Chesebro, B.; and Morgan, T. 1996. Multidrugresistant Human Immunodeficiency Virus Type 1 Strains Resulting from Combination Antiretroviral Therapy. *Journal of Virology* 70(2): 1086–1090.

Jacobsen, H.; Hanggi, M.; Ott, M.; Duncan, I.; Owen, S.; Andreoni, M.; Vella, S.; and Mous, J. 1996. In Vivo Resistance to a Human Immunodeficiency Type-1 Proteinase Inhibitor. *Journal of Infectious Diseases* 173(6): 1379–1387.

Lathrop, R. H., and Pazzani, M. J. 1999. Combinatorial Optimazation in Rapidly Mutating Drug-Resistant Viruses. *Journal of Combinatorial Optimazation.* Forthcoming.

Lange, J. 1995. Triple Combinations: Present and Future. *Journal of AIDS and Human Retrovirology* 10(S1): S77–82.

Musen, M.; Wieckert, K.; Miller, E.; Campbell, K.; and Fagan, L. 1995. Development of a Controlled Medical Terminology: Knowledge Acquisition and Knowledge Representation. *Methods of Information in Medicine* 34(1–2): 85–95.

Musen, M.; Tu, S.; Das, A.; and Shahar, Y. 1996. EON: A Component-Based Approach to Automation of Protocol-Directed Therapy. *Journal of the American Medical Informatics Association* 3(6): 367–388.

O'Brian, W.; Hartigan, P.; Martin, D.; Esinhart, J.; Hill, A.; Benoit, S.; Rubin, M.; Simberkoff, M.; and Hamilton, J. 1996. Changes in Plasma HIV-1 and CD4+ Lymphocyte Counts and the Risk of Progression to AIDS. *New England Journal of Medicine* 334(7): 426–431.

Ohno-Machado, L.; Parra, E.; Henry, S.; Tu, S.; and Musen, M. 1993. AIDS 2: A Decision-Support Tool for Decreasing Physician's Uncertainty Regarding Patient Eligibility for HIV Treatment Protocols. In Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care, 429–433. Philadelphia, Pa.: Hanley and Belfus.

Pazzani, M., and Kibler, D. 1992. The Utility of Prior Knowledge in Inductive Learning. *Machine Learning* 9(2): 54–97.

Pazzani, M.; Iyer, R.; See, D.; Shroeder, E.; and Tilles, J. 1997. CTSHIV: A Knowledge-Based System in the Management of HIV-Infected Patients. Paper presented at the International Conference on Intelligent Information Systems, 8–10 December, Grand Bahamas Island, Bahamas.

Pazzani, M.; See, D.; Shroeder, E.; and Tilles, J. 1997. Application of an Expert System in the Management of HIV-Infected Patients. *Journal of AIDS and Human Retrovirology* 15(5): 356–362.

Ruggiero, C.; Giacomini, M.; Varnier, O. E.; and Gaglio, S. 1994. A Qualitative Process Theory–Based Model of the HIV-1 Virus-Cell Interaction. *Computer Methods and Programs in Biomedicine* 43(3–4): 255–259.

Safran, C.; Rind, D.; Sands, D.; Davis, R.; Wald, J.; and Slack, W. 1996. Development of a Knowledge-Based Electronic Patient Record. *M.D. Computing* 13(1): 46–54.

Sieburg, H. 1994. Methods in the Virtual Wetlab I: Rule-Based Reasoning Driven by Nearest-Neighbor Lattice Dynamics. *AI in Medicine* 6(4): 301–319.

Sobesky, M.; Michelet, C.; Thomas, R.; and LeBeux, P. 1994. Decision-Making System. *Journal of Clinical Computing* 22(1): 20–26.

Sonnenberg, F.; Hagerty, C.; and Kulikowski, C. 1994. An Architecture for Knowledge-Based Construction of Decision Models. *Medical Decision Making* 14(1): 27–39.

Tu, S.; Eriksson, H.; Gennari, J.; Shahar, Y.; and Musen, M. 1995. Ontology-Based Configuration of Problem-Solving Methods and Generation of Knowledge-Acquisition Tools. *AI in Medicine* 7:257–289.

Wong, J.; Hezareh, M.; Gunthard, H.; Havlir, D.; Ignacio, C.; Spina, C.; and Richman, D. 1997. Recovery of Replication-Competent HIV Despite Prolonged Suppression of Plasma Viremia. *Science* 278(5341): 1291–1295.

Xu, L. 1996. An Integrated Rule- and Case-Based Approach to AIDS Initial Assessment. *International Journal of Bio-Medical Computing* 40:197–207.

Richard H. Lathrop is an associate professor of Information and Computer Science at the University of California at Irvine, conducting research in AI and computational molecular biology. He received a Ph.D. in AI from the Massachusetts Institute of Technology in 1990 and is a life mem-

ber of the American Association of Artificial Intelligence. His publications include the cover articles for issues of *Communications of the ACM* and *Journal of Molecular Biology.* He was a founding member of the Board of Directors of the International Society for Computational Biology; is a cofounding scientist of Arris Pharmaceutical Corp.; and is on the Scientific Advisory Board of CombiChem, Inc.

Nicholas R. Stefien is a Ph.D. student in information and computer science at the University of California at Irvine. He is also a senior staff scientist at Raytheon Systems Company. His research and professional interests include computational molecular biology, probabilistic methods of image processing, and hardware-verification methods.

Miriam P. Raphael received her bachelor's degree in information and computer science from the University of California at Irvine (UCI). She worked on several independent undergraduate research projects at UCI, including CTSHIV. She is currently a graduate student at Johns Hopkins University specializing in computational biology.

Sophia Deeds-Rubin holds dual bachelor's degrees, one in information and computer science from the University of California at Irvine and another from Beltsy College of Music (in the former Soviet Union). Her interest in computational molecular biology resulted in several independent undergraduate research projects related to CTSHIV. Currently, she is working as an MTA associate on the Wide Area Augmentation System Program at Raytheon Systems Company in Fullerton, California.

Michael J. Pazzani is chair of the Department of Information and Computer Science at the University of California at Irvine (UCI). He joined UCI in 1988 after receiving a Ph.D. in computer science from the University of California at Los Angeles. His research interests include intelligent software agents, AI, and machine learning. He has done research in expert systems for space and transportation applications and pattern discovery in databases for medical, financial, political, and telecommunications applications. Pazzani is the author of two books on machine learning.

Paul J. Cimoch is director of medical services and founder of the Orange County Center for Special Immunology (OCCSI). Established in Irvine, California, in 1990, OCCSI provides comprehensive primary care and clinical research for immune system disorders. Cimoch presented research at the last eight international conferences on AIDS and is the author of many HIV-related publications. He received his medical degree from the University of Texas Medical Branch in Galveston and completed his residency in internal medicine at the University of Miami School of Medicine and Jackson Memorial Hospital. He is board certified in internal medicine and forensic medicine and is a fellow of the American College of Physicians.

Darryl M. See is an associate professor of medicine at the University of California at Irvine; is the recipient of numerous awards and research grants; is the author of several textbook chapters and dozens of journal articles; has pioneered the use of AI and computer programs in the treatment of AIDS; has one of the largest chronic fatigue clinics in the nation; and has published extensively on enteroviral infections.

Jeremiah G. Tilles M.D., M.A.C.P., is a professor of medicine and chief of the Division of Infectious Diseases at the University of California at Irvine (UCI). He has 97 publications primarily in the field of antiviral agents and has served on the National Institute of Allergy and Infectious Diseases Antiviral Substance Study Group. He received his M.D. from Harvard Medical School and went on to residency in internal medicine at Boston City Hospital (Harvard); a fellowship in infectious diseases under Maxwell Finland, M.D., at the Thorndike Memorial Laboratory; and a special postdoctoral fellowship on Interferon under Alik Issacs at the National Institute for Medical Research (Mill Hill, London). He served five years on the faculty at Harvard Medical School before transferring to UCI.