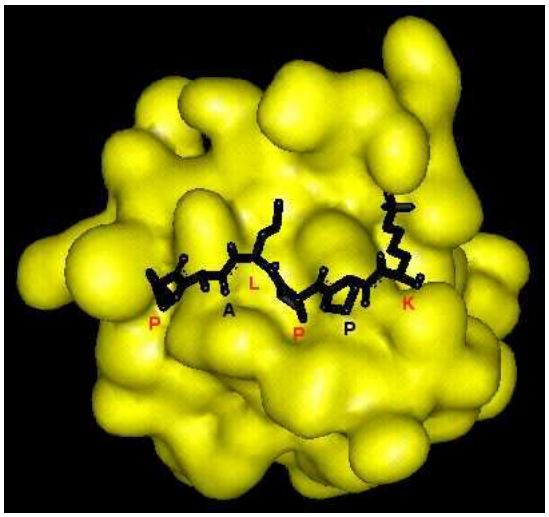
# Applications of Inductive Logic Programming

Applications in Drug Design

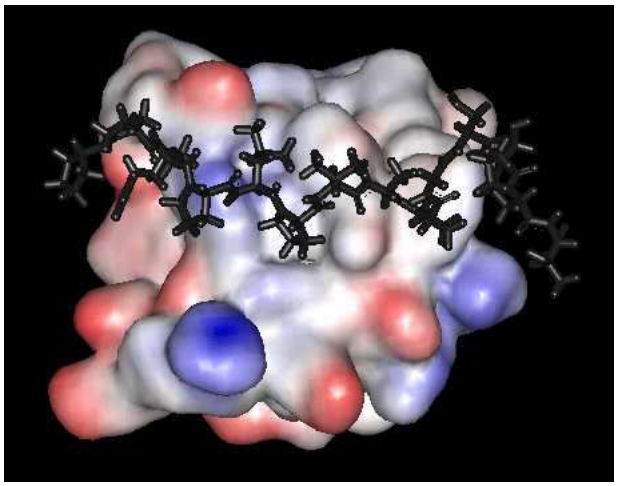
## Background

- A drug is a (typically small, organic) molecule that binds to a larger biomolecule (almost always a protein) within the body.
- The protein may belong to the human body or to an invading organism.
- Binding is *not* usually via ordinary covalent bonds but electrostatic (charged) interactions and hydrophobic interactions.

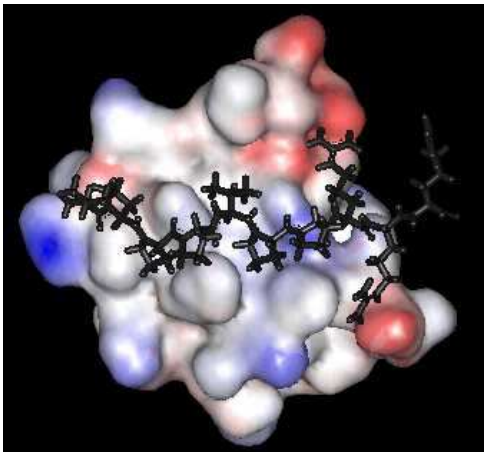
## Crk SH3




## Src SH3



## Grb2 SH3



## Some Examples

- *Selective serotonin reuptake inhibitors:* delay transport of serotonin back into cells. Useful in treating depression.
- *ACE inhibitors:* reduce blood pressure by competitively binding to Angiotensin Converting Enzyme (ACE).
- *Antibiotics:* larger class of drugs. Typically bind to and disrupt behavior of proteins in bacterial cell wall.

## More Background: Proteins

- A protein is a sequence of amino acid "residues" joined by amide bonds. We say "residue" because building an amide bond between two amino acids releases a water molecule.
- 20 common amino acids, represented by single letters. Proteins are commonly 200-300 residues, though they vary from 2-10

## Proteins (Continued)

- (Continued)… (peptides) to over 1000 residues in length. Example: QPPATF.
- The residue sequence is also called the *primary structure*.
- Parts of the sequence then arrange themselves into *alpha helices* (like a phone cord), *beta sheets* (areas that wrap back on

## Proteins (Continued)

- (Continued)… themselves to form a planar sheet. Other regions are called *loop* or *coil*.
- The protein then takes on an overall shape, called the *tertiary structure*. Many of the most interesting proteins are globular (next slide), reflecting the tightness and complexity of the tertiary structure.
- Proteins can even combine or complex with other atoms to form *quarternary structure*.

## Proteins (Continued)

- For our purposes, proteins are interesting because they direct most cell (and hence organism) processes. To modify a process, we'd like to build a molecule to bind to an appropriate site in the appropriate protein.
- Ability to bind is based not only on primary structure but also on the higher levels of protein structure.

## The Drug Design Process (Typical -- Really Varies a lot)

- Identify a target protein such that binding will have the desired effect (may be an agonist or antagonist).
- If we know the 3-dimensional structure of the target site (by X-ray crystallography or NMR spectroscopy), we can design a small molecule that will fit into it.
- **Problem**: We rarely know the 3-D structure

## Drug Design (Continued)

- (Continued)… of the target site. And it is very difficult to predict from the primary structure (amino acid sequence). This is the famous *protein folding* problem.
- But it is possible to test molecules in a test tube to see if they bind to the target site.
- Therefore, what most university researchers and pharmaceutical companies do is to test many compounds in the search for *hits*.

## Drug Design (Continued)

- **Problem 2**: A hit rarely can be used as a drug. The odds are that the molecule either is toxic, has other undesirable side effects, binds a little too loosely (or too tightly), does not diffuse from the gut to bloodstream (so can't be taken by mouth), metabolizes too quickly in the body (breaks down before it can act), metabolizes too slowly (body can't get rid of it eventually), etc.

## Drug Design (Continued)

- So once a few hits are found, the search *begins*. Can we use these hits as *leads* in our search for a drug? We would like to figure out what these hits have in common structurally that is making them active, and make new compounds with these structural features. We'll vary other portions to get around the problems.

## Structure-Activity Prediction (SAR)

- Attempt to predict biological activity of a molecule based on its structure.
- Earliest approach used linear regression.
- Use active and inactive molecules that have a similar basic structure (scaffold). Use features (independent variables) to describe the positions of variation. Class value (dependent variable) is activity level.

## Problems with this Approach

- Could not use very diverse sets of molecules, because could not capture details of molecular structure.
- Could not capture different mechanisms of activity.
- Did not address 3-D issues, and these are crucial for binding.

## Current Practice in Pharmaceutical Industry

- Approach 1: use “2.5-D” representations and decision trees (recursive partitioning).
- Approach 2: use “3-D” representation and special-purpose algorithms.

## Approach 1

- 2.5-D features:
- Basic nitrogen 2 bonds from oxygen.
- 2 Benzene rings 3 bonds apart.
- Recursive Partitioning:
- Can capture different mechanisms.
- Leaf nodes can predict activity level.

## Approach 1 Problems

- Only approximates the 3-D information that is at the heart of ligand-receptor binding.
- Millions of features.

## Approach 2

- Special purpose algorithms estimate the shapes of small molecules, align the possible shapes of the different molecules, and find common “subgraphs”.
- Examples include DISCO by Tripos and Catalyst by MSI.

## Problems with Approach 2

- A molecule can take on several different 3-D shapes, known as *conformations*.
- Doesn't build on lessons from other applications of machine learning and data mining.

## More about Conformations

- A molecule can rotate around single bonds (also a few other changes, such as ring flips).
- Gives rise to an infinite number of different 3-D structures for one molecule.
- Every structure has an energy level.
- Molecules will settle (for a short while) in local minima relative to the energy function.

## Conformations (Continued)

- In a test tube of a given compound, molecules in the test tube will represent a variety of conformations, and each molecule is constantly changing conformation.
- The fraction of molecules in a given conformation in the test tube at a given time is inversely related to the energy level of the

## Conformations (Continued)

- (Continued)… conformation.
- Moreover, the probability (or frequency) of a conformation dies off exponentially with the energy of the conformation. (Because more energy is required to force a molecule into this conformation, and with a lot of energy it may as well expend that energy going to another conformation).

## The Multiple Instance Problem

- If a molecule is *active* (binds to some target protein), it could be because of any of the relatively low energy conformations of the molecules.
- We don't know which conformation is making the molecule active. (We can't isolate a *conformer* because they interconvert so rapidly.)

## Mutiple Instance (Continued)

- If we wish to represent the 3-D structure of a molecule with a feature vector, we must use *multiple* feature vectors (multiple instances), one for each conformer.
- The class value (e.g., *active* or *inactive*) is associated with the example (set of vectors), and not an individual vector (instance) itself.

## Multiple Instance (Continued)

- If an example is *positive* (e.g., active), we only know that at least one of the feature vectors for that example is *positive*.
- If an example is *negative*, then all the vectors for that example are negative.
- This setting requires significant changes to most machine learning algorithms.

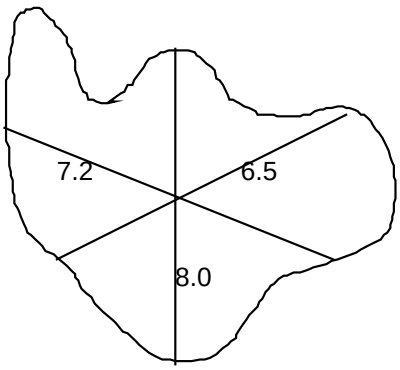
## Generality of the Problem

- The multiple instance problem can also arise within other settings:
- Classifying web sites as interesting.
- Flagging satellite images for inspection.
- But it was first raised by Dietterich, Lathrop, and Lozano-Perez in this very context of drug design (AIJ 1994).

# COMPASS

- COMPASS is a neural network system for structure-activity prediction.
- Represents a conformer by a feature vector of diameter lengths.
- An example (molecule) is represented by a set of vectors, one for each of its conformers.

# COMPASS (Illustration)




7.2 6.5 8.0 Active

## COMPASS Procedure

- Align conformers of different molecules.
- Generate feature vectors.
- Train neural network specifically designed for multiple instance problem on the given feature vectors.
- Possibly repeat above process using the neural net results to direct conformational alignments.

## Limitations

- Must *align* conformers of different molecules. A mistake here can ruin results.
- Assumes *only* shape matters. Makes it difficult to encode locations of interesting atoms or groups, such as basic nitrogens or hydrophobic groups. (We can do this if every molecule has exactly the same number and type of groups of interest).

## Another Alternative: ILP

- Again compute low-energy conformers of the molecules.
- Represent 2-D and 3-D structures of molecules explicitly in background knowledge.
- Define atoms and groups of interest in background knowledge.
- Represent geometry by pairwise distances; define distance computation in background.

## Details of ILP for Drug Design

- Recall we want to learn one or more rules that describe structures predictive of a biological activity.
- We wish to use 3-D conformers of molecules, and learn 3-D structures.
- Such a structure often is called a *pharmacophore*.

ILP for pharmacophore discovery

Represents data points in mathematical logic
    expressive enough to encode 3D structural data
Uses Background Knowledge
    e.g. hydrophobic groups, Euclidean distances
Returns results in logic
    easily translated into English rules

# Pharmacophore Data

**Given:** Positive examples—active molecules
Negative examples—inactives
(optional)
Background knowledge—relevant groups, geometry definitions, etc.

**Find:** A *hypothesis* that, together with the background knowledge, logically implies the data (examples).

In the present work, a *hypothesis* is a *pharmacophore* that is common to all the active molecules (but not possessed by inactives).

# Background Knowledge

- Information about atoms and bonds in the molecules -- include conformer ID as an argument, since positions of atoms may vary from one conformer to another.

- atm(m1,c1,a1,o,3,5.915800,-2.441200,1.799700).
- atm(m1,c1,a2,c,3,0.574700,-2.773300,0.337600).
- atm(m1,c1,a3,s,3,0.408000,-3.511700,-1.314000).

- bond(m1,c1,a1,a2,1).
- bond(m1,c1,a2,a3,1).

# Background knowledge (Continued)

- Definition of distance equivalence

- dist(Molecule,Atom1,Atom2,Dist,Error):-
- number(Error),
- coord(Molecule,Atom1,X1,Y1,Z1),
- coord(Molecule,Atom2,X2,Y2,Z2),
- euc_dist(p(X1,Y1,Z1),p(X2,Y2,Z2),Dist1),
- Diff is Dist1-Dist,
- absolute_value(Diff,E1),
- E1 =< Error.

- euc_dist(p(X1,Y1,Z1),p(X2,Y2,Z2),D):-
- Dsq is (X1-X2)^2+(Y1-Y2)^2+(Z1-Z2)^2,
- D is sqrt(Dsq).

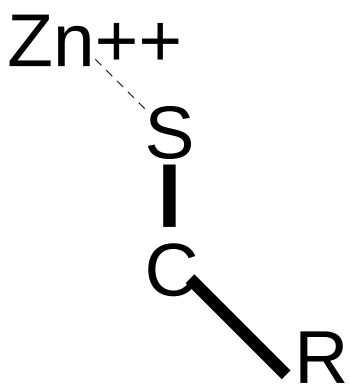
# Background knowledge (Continued)

- Define groups or atom types of chemical significance, and relevance to the biological activity in question. Definitions of hydrophobic groups and zinc binders are somewhat more complex than these.

- hacc(M,C,A):-atm(M,C,A,o,2,_,_,_).
- hacc(M,C,A):-atm(M,C,A,n,ar,_,_,_).

- hdonor(M,C,A):- atm(M,C,A,h,_,_,_,_),
- sym_bond(M,C,A,A1,1),
- atm(M,C,A1,T,_,_,_,_),
- T \= c.

# Background Knowledge (Groups that Bind Zinc)



# Background knowledge (Sample of Groups that Bind Zinc)

```
thiol(M,C,A,s(X1,Y1,Z1),c(X2,Y2,Z2),r(X3,Y3,Z3)):-
        atm(M,C,A,s,_,X1,Y1,Z1),
        atm(M,C,A2,c,_,X2,Y2,Z2),
        sym_bond(M,C,A,A2,1),
        sym_bond(M,C,A2,A3,_),
        A3 \= A,
        atm(M,C,A3,T,_,X3,Y3,Z3),
        T \= h.
```

# The Logical Representation of a Pharmacophore

*Active(X)* if: *has-conformation(X,Conf),*
*has-hydrophobic(X,A),*
*has-hydrophobic(X,B),*
*distance(X,Conf,A,B,3.0,1.0),*
*has-acceptor(X,C),*
*distance(X,Conf,A,C,4.0,1.0),*
*distance(X,Conf,B,C,5.0,1.0).*

This logical clause states that a molecule X is active if it has some conformation Conf, hydrophobic groups A and B, and a hydrogen acceptor C such that the following holds: in conformation Conf of molecule X, the distance between A and B is 3 Angstroms (plus or minus 1), the distance between A and C is 4, and the distance between B and C is 5.

## Details of Specifying the Hypothesis Language

- Easiest to describe in the context of an example.
- To simplify the example, let's assume for the time being that we have just one conformer per molecule. (This assumption also is made in the on-line demo available in ~cs731-1/public/Pharm._

## Specifying this Hypothesis Language in Progol

- *modeh(*,active(+molecule))*: says any hypothesis must have a head literal built from the unary predicate *active*, whose sole argument is a variable, and this variable will always be bound to a term denoting a *molecule*.
- *molecule* is a type. It can be defined intensionally or extensionally. For example

## Modes and Types (Continued)

- (Continued)… we might use facts of the form
- *molecule(m1).*
- *molecule(m2).*
- Or we might write
- *molecule(X):- name(X,[109|T]),*
- *name(N,T),*
- *integer(T).*

## Modes (Continued)

- To specify the forms of literals in the body, we use *modeb* rather than *modeh.* For example, if we wish to use only (any number of) literals built from the predicates *has_acceptor* and *dist*, then we would have the following modes:
- *modeb(*,hacc(+molecule,-atom))*
- *modeb(*,dist(+molecule,+atom,*
- *+atom,#dist,#error)).*

## Finer Points on Modes

- A "+" designates an input variable (must be bound when this literal is reached), a "-" designates an output variable (need not be bound), and a "#" designates a constant.
- The first argument in the mode is a recall number. If it is *n*, then for any binding of the input variables, a hypothesis can contain only *n* literals. A * means there is no limit on the number of literals.

## Use of Modes

- Modes are used in the construction of the bottom clause (the saturation of a positive example for directing the search).
- Easiest to see with an example. Assume we have the background knowledge already described, and we wish to learn a clause built from predicates *hacc*, *hdonor*, *hydrophobic*, and *dist*.

## Building a Bottom Clause

- Suppose we choose *active(m1)* as our seed example. The bottom clause must have a head of the form *active(X)* because of our head mode. And *X* will be bound to *m1*.
- We now add all the literals we can to the body of this bottom clause that (1) agree with the body modes, and (2) still allow the seed example to be proven.

## Example (Continued)

- If *m1* has two hydrogen acceptors, we might next add literals such as *hacc*(*X,A)* and *hacc(X,B)*.
- If these two hydrogen acceptors are 5.4 Angstroms apart, we can now add a literal *dist(X,A,B,5.4,1.0)*. (I am assuming we have set 1.0 as our error tolerance with a type statement of the form *error(1.0)*.)

## Example (Continued)

- We would continue adding literals built from *hacc*, *hdonor*, and *hydrophobic* as long as there are other such atoms/groups in *m1* according to the background knowledge. For every pair of these, we will add a corresponding *dist* predicate.
- At the end, we might have a bottom clause such as the following, which is the most

## Example (Continued)

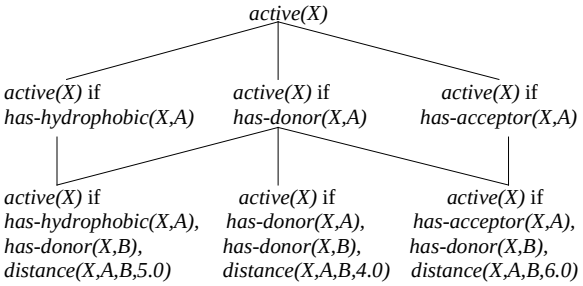
- (Continued)… specific clause (under subsumption) that agrees with the modes and allows *active(m1)* to be proven.
- *active(X):- hacc(X,A), hacc(X,B),*
- *dist(X,A,B,5.4,1.0),*
- *hdonor(X,C),*
- *dist(X,A,C,3.0,1.0),*
- *dist(X,B,C,4.3,1.0),...*

# The Top-Down Search

- Having constructed a large bottom clause for the seed example, Progol will now search the portion of the subsumption lattice above this bottom clause.
- Progol searches this lattice top-down, starting with *active(X)*, in a branch-and-bound (or A*) style, seeking to maximize a utility function (minimize cost function).

# Lattice for the Given Hypothesis Language

## Lattice of Hypotheses




etc.

## Summary of Progol Algorithm

- Select a positive example as a seed
- Use the seed to eliminate portions of the lattice
- Find more general rule in lattice
- Return all hypotheses that are consistent with the data

## Putting the Conformers Back In

- For simplicity, we omitted the problem of multiple conformers for awhile.
- For multiple conformers, we add one other modeb statement, to incorporate literals of the form *conformation(C).*
- The body of the bottom clause then will be disconnected, with one component per conformer.

# A Bottom Clause with Multiple Conformations

*Active(X)* if: *has-conformation(X,Conf),*
*has-hydrophobic(X,A),*
*has-hydrophobic(X,B),*
*distance(X,Conf,A,B,3.0,1.0),*
*has-acceptor(X,C),*
*distance(X,Conf,A,C,4.0,1.0),*
*distance(X,Conf,B,C,5.0,1.0),…*
*has-conformation(X,Conf2),*
*distance(X,Conf2,A,B,6.0,1.0),*
*distance(X,Conf2,A,C,8.0,1.0),*
*distance(X,Conf2,B,C,7.0,1.0),…*

# Conformational model

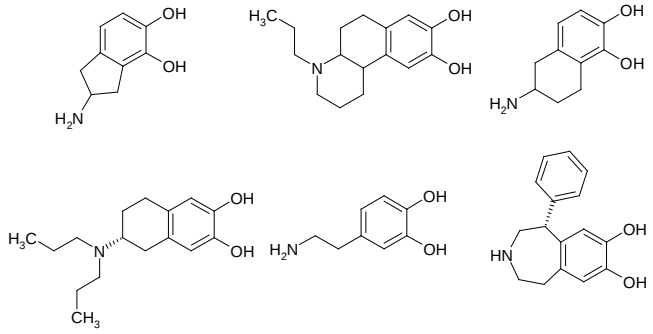
- Conformational flexibility modelled as multiple conformations:
  - Sybyl randomsearch
  - Catalyst

# Pharmacophore description

- Atom and site centred
  - Hydrogen bond donor
  - Hydrogen bond acceptor
  - Hydrophobe
  - Site points (limited at present)
  - User definable
- Distance based

# Example 1: Dopamine agonists

- Agonists taken from Martin data set on QSAR society web pages
- Examples (5-50 conformations/molecule)

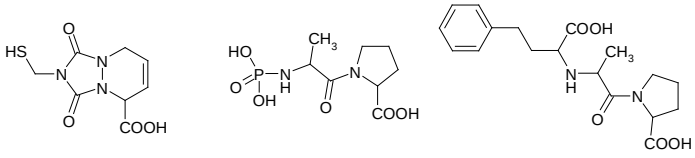


## Pharmacophore identified

- Molecule A has the desired activity if:
- in conformation B molecule A contains a hydrogen acceptor at C, and
- in conformation B molecule A contains a basic nitrogen group at D, and
- the distance between C and D is 7.05966 +/- 0.75 Angstroms, and
- in conformation B molecule A contains a hydrogen acceptor at E, and
- the distance between C and E is 2.80871 +/- 0.75 Angstroms, and
- the distance between D and E is 6.36846 +/- 0.75 Angstroms, and
- in conformation B molecule A contains a hydrophobic group at F, and
- the distance between C and F is 2.68136 +/- 0.75 Angstroms, and
- the distance between D and F is 4.80399 +/- 0.75 Angstroms, and
- the distance between E and F is 2.74602 +/- 0.75 Angstroms.

# Example II: ACE inhibitors

- 28 angiotensin converting enzyme inhibitors taken from literature
  - D. Mayer *et al.*, J. Comput.-Aided Mol. Design, **1**, 3-16, (1987)



# Experiment 1

- Attempt to identify pharmacophore using original Mayer *et al*. Data (final conformations).
- Initial failed attempt traced to “bugs” in background knowledge definition.
- 4 pharmacophores found with corrected code (variations on common theme)

# ACE pharmacophore

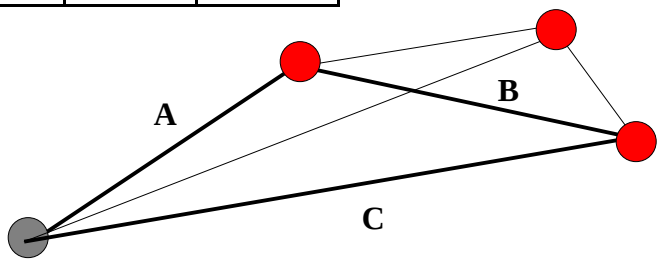
- Molecule A is an ACE inhibitor if:
- molecule A contains a zinc-site B,
- molecule A contains a hydrogen acceptor C,
- the distance between B and C is 7.899 +/- 0.750 A,
- molecule A contains a hydrogen acceptor D,
- the distance between B and D is 8.475 +/- 0.750 A,
- the distance between C and D is 2.133 +/- 0.750 A,
- molecule A contains a hydrogen acceptor E,
- the distance between B and E is 4.891 +/- 0.750 A,
- the distance between C and E is 3.114 +/- 0.750 A,
- the distance between D and E is 3.753 +/- 0.750 A.

# Pharmacophore discovered

| Distance | Progol | Mayer |
|---|---|---|
| A | 4.9 | 5.0 |
| B | 3.8 | 3.8 |
| C | 8.5 | 8.6 |

Zinc site

H-bond acceptor

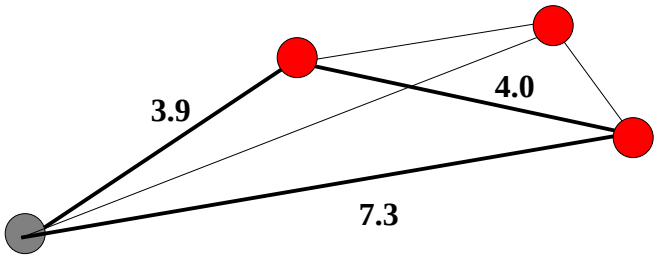



# Experiment 2

- Definition of “zinc ligand” added to background knowledge
  - based on crystallographic data
- Multiple conformations
  - Sybyl RandomSearch

# Experiment 2

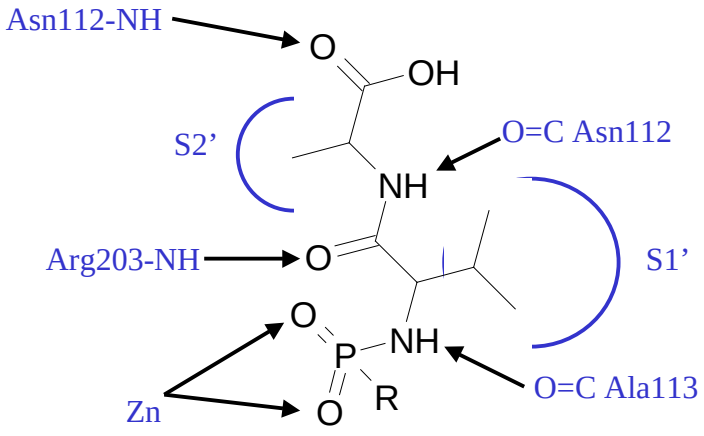
- Original pharmacophore rediscovered plus one other
  - different zinc ligand position
  - similar to alternative proposed by Ciba-Geigy




# Example III: Thermolysin inhibitors

- 10 inhibitors for which crystallographic data is available in PDB
- Conformationally challenging molecules
- Experimentally observed superposition

# Key binding site interactions




# Interactions made by inhibitors

| Interaction | 1HYT | 1THL | 1TLP | 1TMN | 2TMN | 4TLN | 4TMN | 5TLN | 5TMN | 6TMN |
|---|---|---|---|---|---|---|---|---|---|---|
| Asn112-NH | | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ |
| S2’ | | ϑ | ϑ | ϑ | | | ϑ | ϑ | ϑ | ϑ |
| Asn112 C=O | | ϑ | ϑ | ϑ | | | ϑ | ϑ | ϑ | ϑ |
| Arg 203 NH | ϑ | ϑ | ϑ | ϑ | ϑ | | ϑ | ϑ | ϑ | ϑ |
| S1’ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ |
| Ala113-C=O | | | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | |
| Zn | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ | ϑ |

## Pharmacophore Identification

- Structures considered 1HYT 1THL 1TLP 1TMN 2TMN 4TLN 4TMN 5TLN 5TMN 6TMN
- Conformational analysis using “Best” conformer generation in Catalyst
- 98-251 conformations/molecule

## Thermolysin Results

- 10 5-point pharmacophore identified, falling into 2 groups (7/10 molecules)
  - 3 “acceptors”, 1 hydrophobe, 1 donor
  - 4 “acceptors, 1 donor
- Common core of Zn ligands, Arg203 and Asn112 interactions identified
- Correct assignments of functional groups
- Correct geometry to 1 Angstrom tolerance

## Thermolysin results

- Increasing tolerance to 1.5Angstroms finds common 6-point pharmacophore including one extra interaction

## Ongoing ILP developments (pharmacophores)

- Continue to extend method validation
- Extending to combinatorial mixtures
- Quantitative models
- Mixing different datatypes in background knowledge
- Developing graphical front-end

## Ongoing Developments

- Analysis of Combinatorial Chemistry data sets
- Parallel processing on Beowulf architecture to speed week+ long searches.
- Making ILP systems more “collaborative” -- proposing experiments, justifying hypothesis, clustering and summarizing hypotheses, pointing out anomalies in data,