Efficient Perception of Proteins and Nucleic Acids from Atomic Connectivity

Roger Sayle, Ph.D.

NextMove Software,

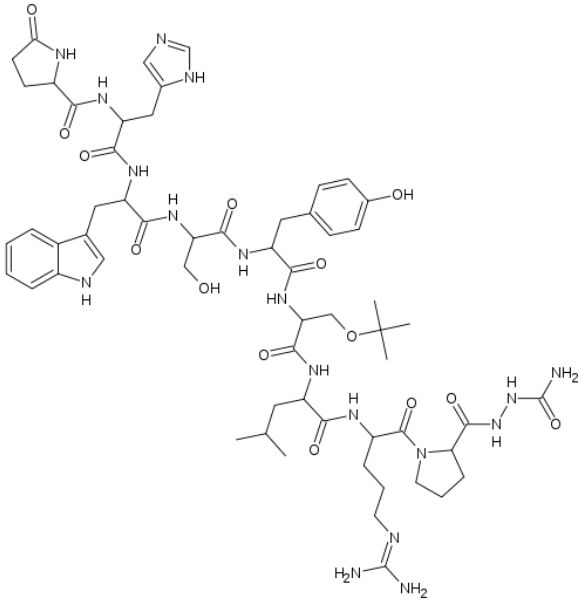
Cambridge, UK

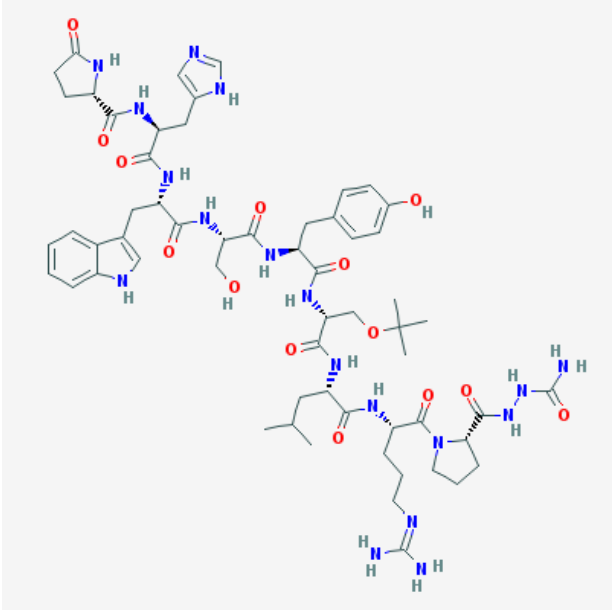
ACS National Meeting, San Diego, USA 26th March 2012

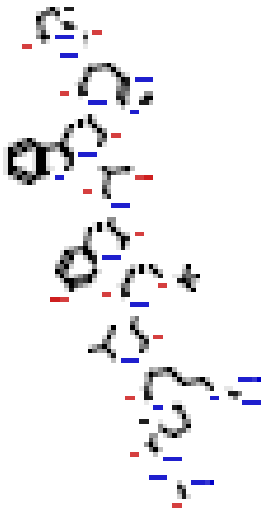
MOTIVATION

- Chemical Structure Normalization/Tautomers.
- IMI Open PHACTS suggest using FDA rules.
- FDA 2007 guidelines on structure registration contains a section on peptides.
- Bridging the gap between cheminformatics (small molecules) and bioinformatics (DNA and protein sequences), peptides pose some interesting technical challenges.

DEPICTION OF PEPTIDES







Goserelin

$C_{59}H_{84}N_{18}O_{14}$

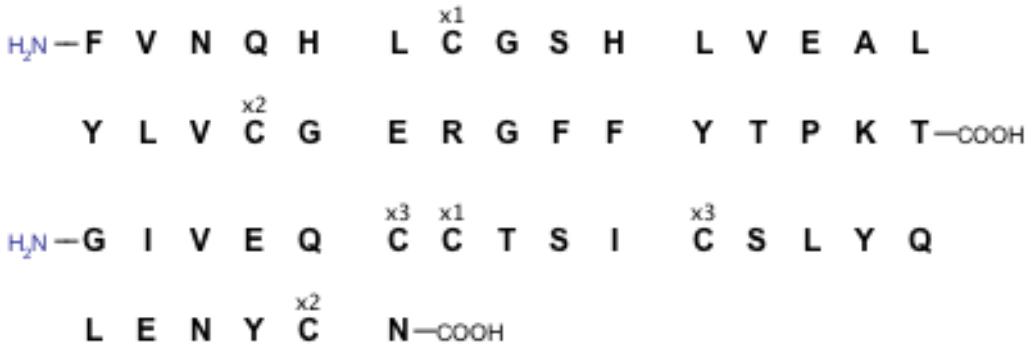
Pyro— Glu— His— Trp— Ser—

Tyr— D— Ser(Bu')— Leu—

Arg— Pro— Azgly— NH_2

END RESULT

CC[C@H](C)[C@H]1C(=O)N[C@H]2CSSC[C@@H](C(=O)N[C@@H](CSSC[C@@H](C(=O)NCC(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@@H](CSSC[C@H](NC(=O)[C@@H](NC(=O)[C@@H](NC(=O)[C@@H](NC(=O)[C@@H](NC(=O)[C@@H](NC(=O)[C@@H](NC2=O)CO)CC(C)C)CC3=CC=C(C=C3)O)CCC(=O)N)CC(C)C)CCC(=O)O)CC(=O)N)CC4=CC=C(C=C4)O)C(=O)N[C@@H](CC(=O)N)C(=O)O)C(=O)NCC(=O)N[C@@H](CCC(=O)O)C(=O)N[C@@H](CCCNC(=N)N)C(=O)NCC(=O)N[C@@H](CC5=CC=CC=C5)C(=O)N[C@@H](CC6=CC=CC=C6)C(=O)N[C@@H](CC7=CC=C(C=C7)O)C(=O)N[C@@H]([C@@H](C)O)C(=O)N8CCC[C@H]8C(=O)N[C@@H](CCCCN)C(=O)N[C@@H]([C@@H](C)O)C(=O)O)C(C)C)CC(C)C)CC9=CC=C(C=C9)O)CC(C)C)C)CCC(=O)O)C(C)C)CC(C)C)CC2=CN=CN2)CO)NC(=O)[C@H](CC(C)C)NC(=O)[C@H](CC2=CN=CN2)NC(=O)[C@H](CCC(=O)N)NC(=O)[C@H](CC(=O)N)NC(=O)[C@H](C(C)C)NC(=O)[C@H](CC2=CC=CC=C2)N)C(=O)N[C@H](C(=O)N[C@H](C(=O)N1)CO)[C@@H](C)O)NC(=O)[C@H](CCC(=O)N)NC(=O)[C@H](CCC(=O)O)NC(=O)[C@H](C(C)C)NC(=O)[C@H]([C@@H](C)CC)NC(=O)CN Human Insulin



FILE FORMAT CONVERSION

- Information loss from MDL/SMILES to PDB/MMD

  Bond orders, Formal charges, Isotopes

- Information loss from PDB/MMD to MDL/SMILES

  Residue information, Occupancy, B-factors

THE CHALLENGE

- Many structural biology and computational chemistry applications assume standard atom and residue naming, some even ordering.
- In OpenEye's OEChem toolkit, recognizing PDB residues is done in OEPercieveResidues, and in OpenBabel it's done in OBChainsParsers.
- This talk describes aspects of the advanced pattern matching algorithms that are used.

THE INPUT

- The input is a minimal connection table.
- Only the graph connectivity is used, allowing molecules without bond orders to be handled.
- Only heavy atoms are required, explicit and implicit hydrogen counts aren't used.
- Likewise, any formal charges, hybridization and 3D geometry are also ignored.
- Typical sources include SMILES and XYZ files.

THE OUTPUT (PDB FILE FORMAT)

```
ATOM    261  N   GLY    37      20.172  17.730   6.217  1.00  8.48
ATOM    262  CA  GLY    37      21.452  16.969   6.513  1.00  9.20
ATOM    263  C   GLY    37      21.143  15.478   6.427  1.00 10.41
ATOM    264  O   GLY    37      20.138  15.023   5.878  1.00 12.06
ATOM    265  N   ALA    38      22.055  14.701   7.032  1.00  9.24
ATOM    266  CA  ALA    38      22.019  13.242   7.020  1.00  9.24
ATOM    267  C   ALA    38      21.944  12.628   8.396  1.00  9.60
ATOM    268  O   ALA    38      21.869  11.387   8.435  1.00 13.65
ATOM    269  CB  ALA    38      23.246  12.697   6.275  1.00 10.43
ATOM    270  N   THR    39      21.894  13.435   9.436  1.00  8.70
ATOM    271  CA  THR    39      21.936  12.911  10.809  1.00  9.46
ATOM    272  C   THR    39      20.615  13.191  11.521  1.00  8.32
ATOM    273  O   THR    39      20.357  14.317  11.948  1.00  9.89
ATOM    274  CB  THR    39      23.131  13.601  11.593  1.00 10.72
ATOM    275  OG1 THR    39      24.284  13.401  10.709  1.00 11.66
ATOM    276  CG2 THR    39      23.340  12.935  12.962  1.00 11.81
ATOM    277  N   CYS    40      19.827  12.110  11.642  1.00  7.64
ATOM    278  CA  CYS    40      18.504  12.312  12.298  1.00  8.05
ATOM    279  C   CYS    40      18.684  12.451  13.784  1.00  7.63
ATOM    280  O   CYS    40      19.533  11.718  14.362  1.00  9.64
ATOM    281  CB  CYS    40      17.582  11.117  11.996  1.00  7.80
ATOM    282  SG  CYS    40      17.199  10.929  10.237  1.00  7.30
```

CHAINS ALGORITHM OVERVIEW

- Identify biopolymer backbones.
- Recognize biopolymer monomer sidechains.

CHAINS ALGORITHM OVERVIEW

- Determine connected components.
- Identify “Hetero” atoms (water, solvent, metals, ions, cofactors and common ligands; ATP, NAD, HEM).
- Identify biopolymer backbones (peptide, DNA, RNA).
- Traverse and number biopolymer backbone residues.
- Recognize biopolymer monomer sidechains (potentially substituted functional groups).
- Annotate hydrogen atoms (if present).
- Sort atoms into PDB order.
- Assign consecutive serial numbers (unique names).

POLYMER MATCHING

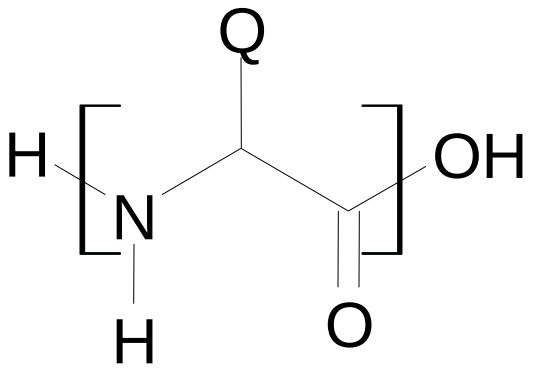
- Matching of linear, cyclic and dendrimeric polymers and copolymers can be efficiently implemented by graph relaxation algorithms.

- Each atom records a set of possible template equivalences represented as a bit vector.

- These bit vectors are iteratively refined using the bit vectors of neighboring atoms.

PEPTIDES AND PROTEINS

- Example of a copolymer (substituted backbone).



- Complications from glycine and proline.
- N-terminal acetyl and formyl.
- C-terminal aldehyde and amide.

PROTEIN BACKBONE DEFINITION

- Assign initial constraints based upon atomic number and heavy atom degree.

BitN	Nitrogen	2 neighbors
BitCA	Carbon	3 neighbors
BitC	Carbon	3 neighbors
BitO	Oxygen	1 neighbor

PROTEIN BACKBONE DEFINITION

- Perform iterative relaxation at each atom using its neighbor's bit masks.

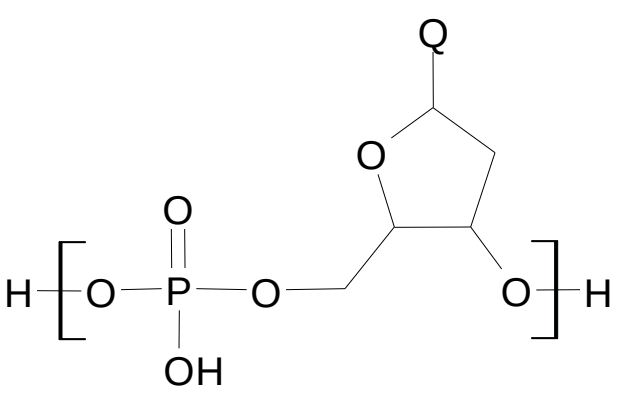
BitN:	BitCA	BitC	
BitCA:	BitN	BitC	AnyC
BitC:	BitCA	BitN	BitO
BitO:	BitC		

CURRENT PROTEIN DEFINITIONS

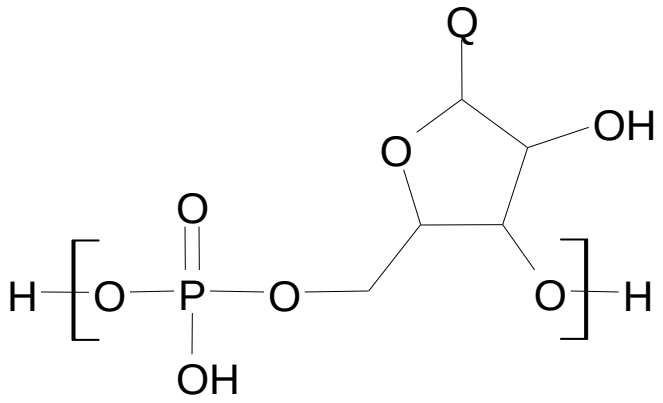
- Five types of nitrogen (±proline, ±terminal & amide)
- Two types of alpha carbon (regular vs. GLY)
- Four types of carbonyl carbon
- Two types of oxygen (O and OXT)

- Currently 13 atom types (bits).

NUCLEIC ACIDS



DNA



RNA

- Complications include termination & methylation.
- Currently 19 atom types (bits).

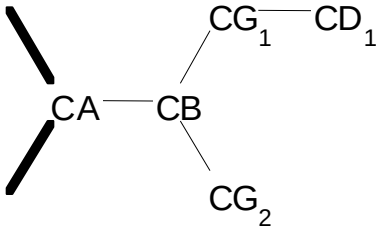
FUNCTIONAL GROUP MATCHING

- Given an known attachment atom/bond determine whether the affixed molecular fragment is a member of a dictionary of known functional groups or sidechains.

- The standard “default” strategy is to treat the dictionary as a list of SMARTS patterns and loop over each one until a hit is identified.
- This is O(n) in the size of the dictionary.

RECOGNIZING MONOMER SIDECHAINS

- Avoid backtracking by preprocessing monomer sets.



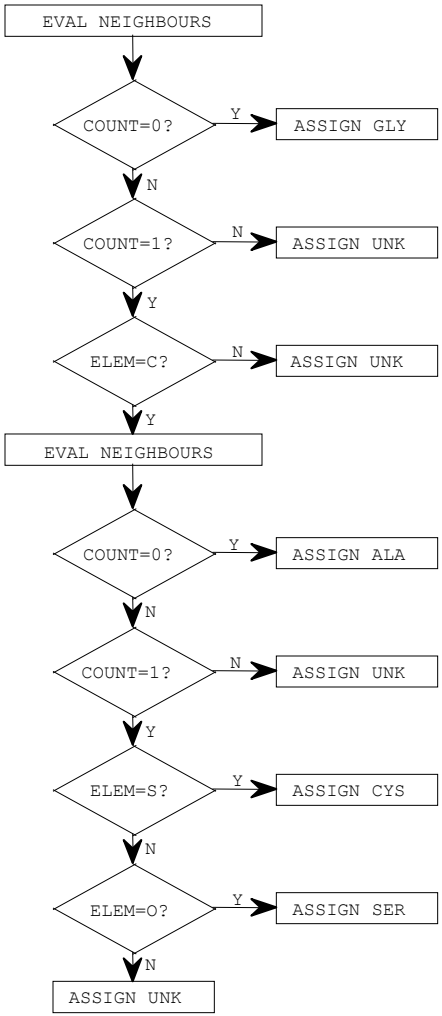
- Enumerate all possible depth first graph traversals

C1,C3,C2,C1,C1 => ILE: CA,CB,CG1,CD1,CG2
C1,C3,C1,C2,C1 => ILE: CA,CB,CG2,CG1,CD1

RECOGNIZING MONOMER SIDECHAINS

- The sequence of atomic numbers and the heavy degree (or fanout) of each vertex uniquely identifies the sidechain.
- Conceptually, the candidate sidechain is traversed once and identified by looking up the “traversal” in a “dictionary”.
- Encode the traversal dictionary as a prefix-closed “tries” to efficiently match arbitrary dictionaries.

EXAMPLE PEPTIDE FLOW CHART



GRAPH MATCHING DURING TRAVERSAL

- Instead of initially traversing the entire sidechain to be recognized and then looking it up in the dictionary, we can perform the look-up as we traverse our molecule.
- This has the benefit that we can fail fast and run-time is bounded by the number of atoms in our largest monomer.
- In fact, the algorithm's performance is independent of the number of monomer definitions.

OECHEM IMPLEMENTATION

- In OEChem, a way-ahead-of-time “compiler” processes an input set of sidechain definitions and outputs them as a bytecode program.
- These bytecodes are efficiently interpreted at run-time inside OEPerceiveResidues.
- Separate sidechain “programs” are used for proteins and nucleic acids.
- An early implementation in OELib and OpenBabel builds these bytecodes just-in-time.

RESULTING BYTECODE

```
static const OEByteCode peptide_bc[301] = {
  { OEOpCode::EVAL,                      0,     1,     0 }, //     0
  { OEOpCode::COUNT,                     0,    61,    60 }, //     1
  { OEOpCode::ELEM,                      6,     3,    -1 }, //     2
  { OEOpCode::EVAL,                      0,     4,     0 }, //     3
  { OEOpCode::COUNT,                     0,     5,     6 }, //     4
  { OEOpCode::ASSIGN, BCNAM('A','L','A'),     2,     0 }, //     5
  { OEOpCode::COUNT,                     1,     7,    77 }, //     6
  { OEOpCode::ELEM,                      6,     8,    41 }, //     7
  { OEOpCode::EVAL,                      0,     9,     0 }, //     8
  { OEOpCode::COUNT,                     0,   174,   173 }, //     9
  { OEOpCode::ELEM,                      7,    11,    17 }, //    10
  { OEOpCode::EVAL,                      0,    12,     0 }, //    11
  { OEOpCode::COUNT,                     0,    13,   229 }, //    12
  { OEOpCode::ELEM,                      8,    14,    -1 }, //    13
  { OEOpCode::EVAL,                      0,    15,     0 }, //    14
  { OEOpCode::COUNT,                     0,    16,    -1 }, //    15
  { OEOpCode::ASSIGN, BCNAM('A','S','N'),     5,     0 }, //    16
```

SUPPORTED PEPTIDE SIDECHAINS

- The twenty standard amino acids (D and L- forms).
  - ALA, ASN, ASP, ARG, CYS, GLN, GLU, GLY, HIS, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR and VAL.
- Fourteen non-standard amino acids.
  - ABA, CGU, CME, CSD, HYP, LYZ, MEN, MLY, MSE, NLE, NVA, ORN, PCA, PTR, SEP and TPO.
- Two N-terminal modifications.
  - Acetyl (ACE) and Formyl (FOR).
- One C-terminal modification residue.
  - Amide (NH2).

SUPPORTED NUCLEIC SIDECHAINS

- Four standard DNA bases.
  - DA, DC, DG, DT.
- Four standard RNA bases.
  - A, C, G, U.
- Seven non-standard RNA bases.
  - 1MA, 2MG, 5MC, 7MG, M2G, PSU, YG.

OEPERCEIVERESIDUES PERFORMANCE

- As measured on a 2.2GHz AMD Opteron.

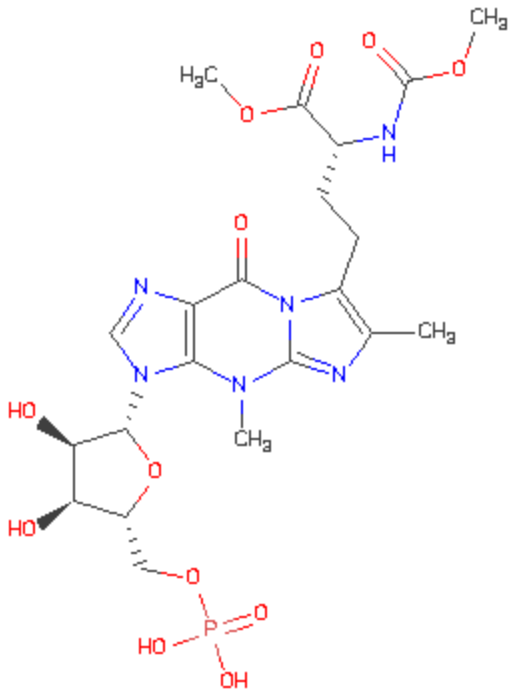
Code	Atoms	Mons	Timings	Average
1CRN	327	46	15s/10K	1.5 ms
4TNA	1,656	76	98s/10K	9.8 ms
1GD1	10,984	1336	185s/1K	185 ms

- Previous work by Siani, Weininger and Blaney (JCICS 1994) report taking 8.2s on an R3000 SGI to identify insulin (51aa) using SMARTS matching.

BYTECODE SIZE ISSUES

- Using all possible traversal permutations works well for the standard amino and nucleic acid sidechains.
- To match the 34 supported amino acid sidechains required a “program” of only 521 bytecodes, and the 11 supported nucleic acid bases required 3,488 byte codes.
- Then came wybutosine (PDB residue YG) which had pathological branching requiring 91,974 byte codes due to combinatorics.

WYBUTOSINE



BYTECODE SIZE ISSUES

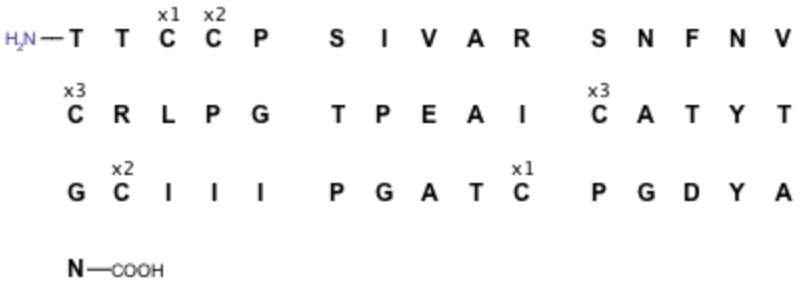
- The solution was to perform a local sorting step when pushing neighboring atoms on to the stack.
- Sort uses atomic number and heavy degree.
- No performance impact in OEChem as we were already sorting for permutation stability.

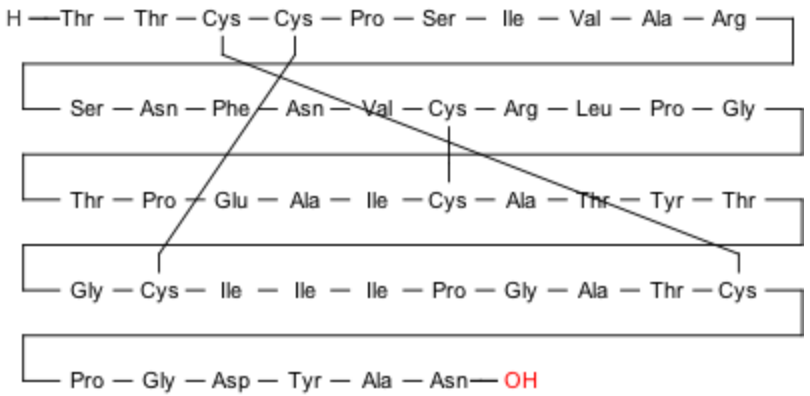
Dictionary	Before	After
Peptide	521	301
Nucleic	3,488	326
Nucleic+YG	95,391	649

IDENTIFYING PEPTIDES

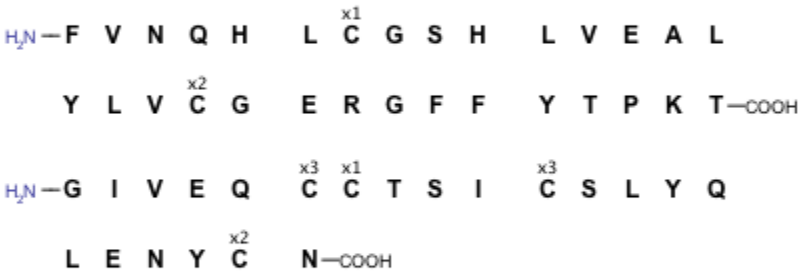
- To a first approximation, use OpenBabel to convert the SMILES to a PDB file, and if it only contains ATOMs (no HETATMs) it's a peptide.
- A better definition is to use an explicit list of acceptable amino (and nucleic) acid residues and check for the presence of these.
- Crosslinking is easily retrieved as bonds between residues, excepting standard eupeptide links.

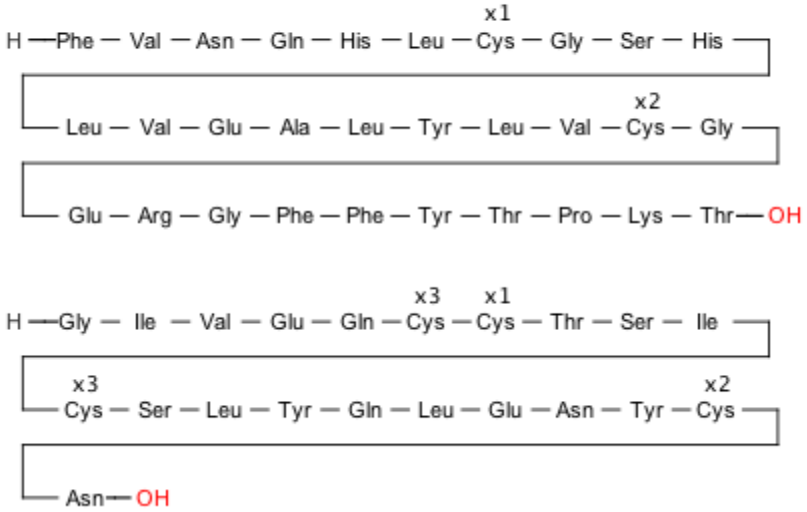
CRAMBIN (FDA & IUPAC STYLES)





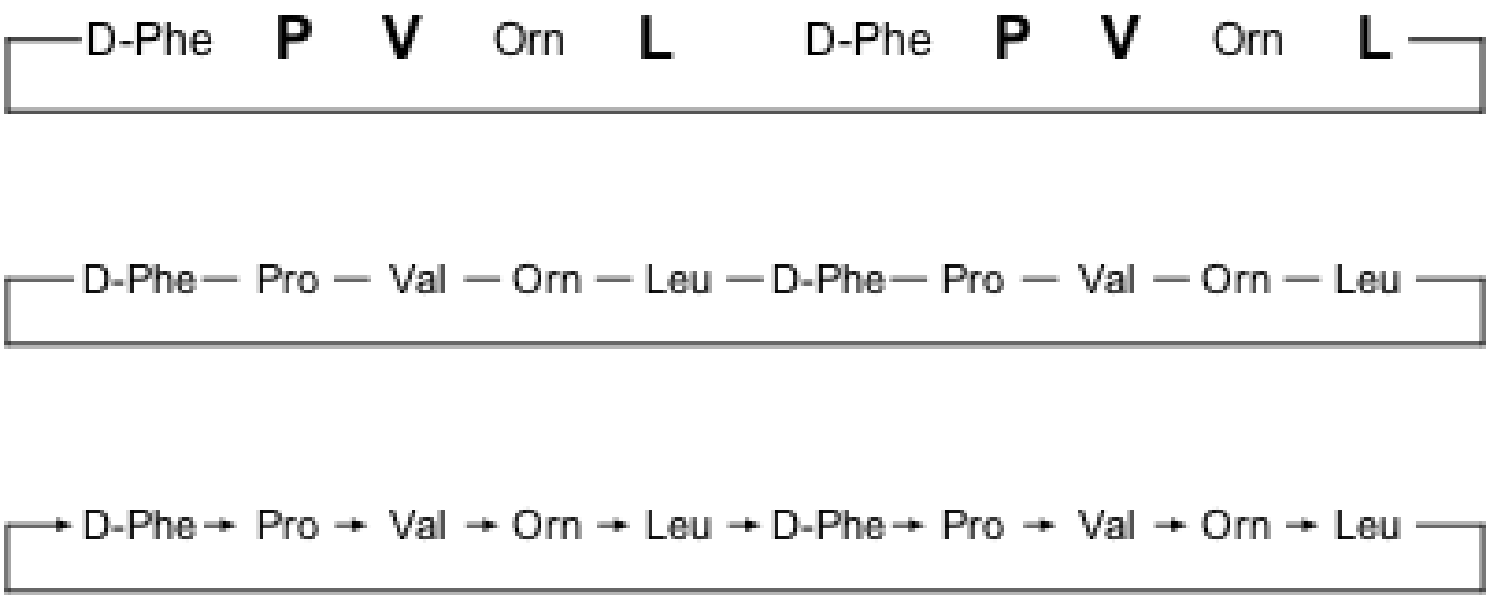
HUMAN INSULIN (MULTIPLE CHAINS)





OXYTOCIN (DISULFIDE & AMIDE)

GRAMICIDIN S (CYCLIC PEPTIDES)



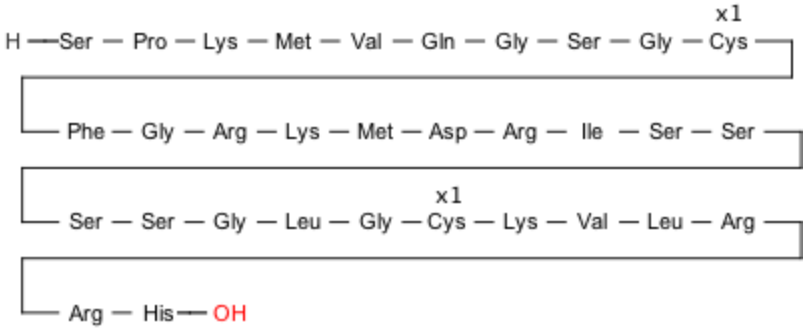
LULIBERIN (TERMINAL OPTIONS)

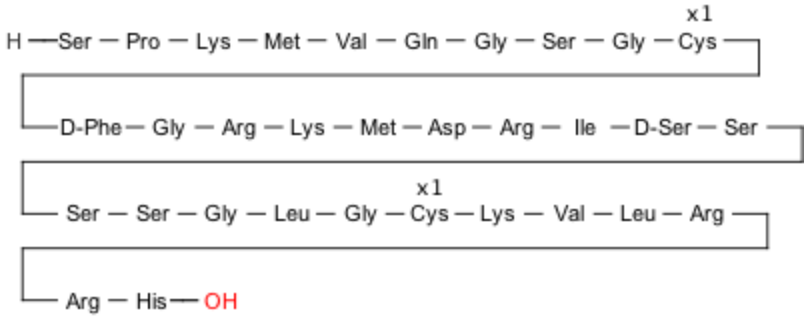
Pyr — His — Trp — Ser — Tyr — Gly — Leu — Arg — Pro — Gly— NH_2

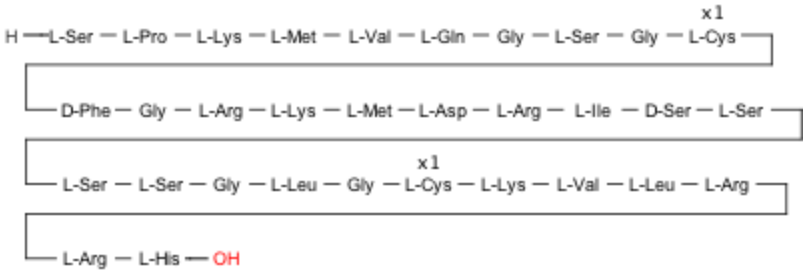
H —Pyr — His — Trp — Ser — Tyr — Gly — Leu — Arg — Pro — Gly— NH_2

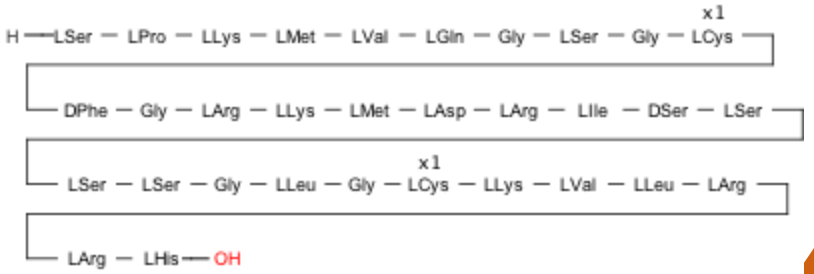
H_2N —Pyr — His — Trp — Ser — Tyr — Gly — Leu — Arg — Pro — Gly—$CONH_2$

NESIRITIDE (STEREO OPTIONS)









DATABASE STATISTICS

- Unfortunately, of 250251 compounds in the NCI Aug 2000 database, only 758 (0.3%) are peptide or peptide-like and only 86 (0.03%) are oligonucleotide sequences.
- Of the 32,274,581 compounds in PubChem[February 2012], only 251,866 (0.78%) and only 1075 (0.00%) are oligonucleotide sequences.
- Alas the SMARTS pattern “NCC(=O)NCC(=O)NCC=O” for a tripeptide only matches 391,502 (1.2%), an upper bound on the prevalence of peptides.

FUTURE WORK

- Analysis of non-standard AA usage in drugs.
- Additional non-standard sidechains (SAR).
- Alloisoleucine (IIL) and allothreonine (ALO).
- Glycoinformatics (post-translational sugar chains).
- Matching improvements (compilation/profiling).

CONCLUSIONS

- Peptides and biologics don't need their own file formats if their sequence can be perceived.
- Generating prettier 2D depictions is simple.
- Peptide informatics (and glycoinfomatics) potentially offer great opportunities for research and innovation.
- It's always great to find a new use for an old algorithm.

ACKNOWLEDGEMENTS

- OpenEye Scientific Software.
- AstraZeneca R&D.
- Vertex Pharmaceuticals.
- European Bioinformatics Institute.
- RasMol's user community.

- Thank you for your time.