HELM: Setting the standard for biomolecular information exchange

CICAG – What would Dalton do now?

Claire Bellamy – HELM project manager

22nd June 2017

Contents

- Why is biomolecular representation an issue?
- What is HELM?
- The HELM ecosystem
- Key areas still to do

Why is biomolecular representation an issue?

And why talk about it here?

Why Now?

Best selling pharmaceuticals 2015

Pistoia Alliance

Rank	Drug	Trade name	Type	Main indications	Company	**Sales** (USD millions/year)	**Δ vs 2014**
1	Adalimumab	Humira	**Biologic**	Rheumatoid arthritis	AbbVie Inc.	14,012	1,469
2	Ledipasvir/sofosbuvir	Harvoni	**Small molecule**	Hepatitis C	Gilead Sciences	13,864	11,737
3	Etanercept	Enbrel	**Biologic**	Rheumatoid arthritis	Amgen Pfizer	8,697	4,009
4	Infliximab	Remicade	**Biologic**	Crohn's Disease Rheumatoid Arthritis	Johnson & Johnson	8,355	1,487
5	Rituximab	Mabthera Rituxan	**Biologic**	Lymphoma Leukemia Autoimmune disorders	Roche	7,115	1,456
6	Insulin glargine	Lantus	**Biologic**	Diabetes mellitus	Sanofi	7,029	51
7	Bevacizumab	Avastin	**Biologic**	Metastatic cancers	Roche	6,751	270
8	Trastuzumab	Herceptin	**Biologic**	Breast cancer	Roche	6,603	265
9	Lenalidomide	Revlimid	**Small molecule**	Multiple myeloma Myelodysplastic syndromes	Celgene	5,801	821
10	Sofosbuvir	Sovaldi	**Small molecule**	Hepatitis C	Gilead Sciences	5,276	(5,007

Source Wikipedia

So what is the problem?

- Small molecule representation is well understood and has been used for many years.
- Although there are still areas to talk about the fundamentals are established.
  - Most compounds pharma can generally be handled
- Biologics is a long way behind.
  - Many important substances cannot be represented fully with current tools

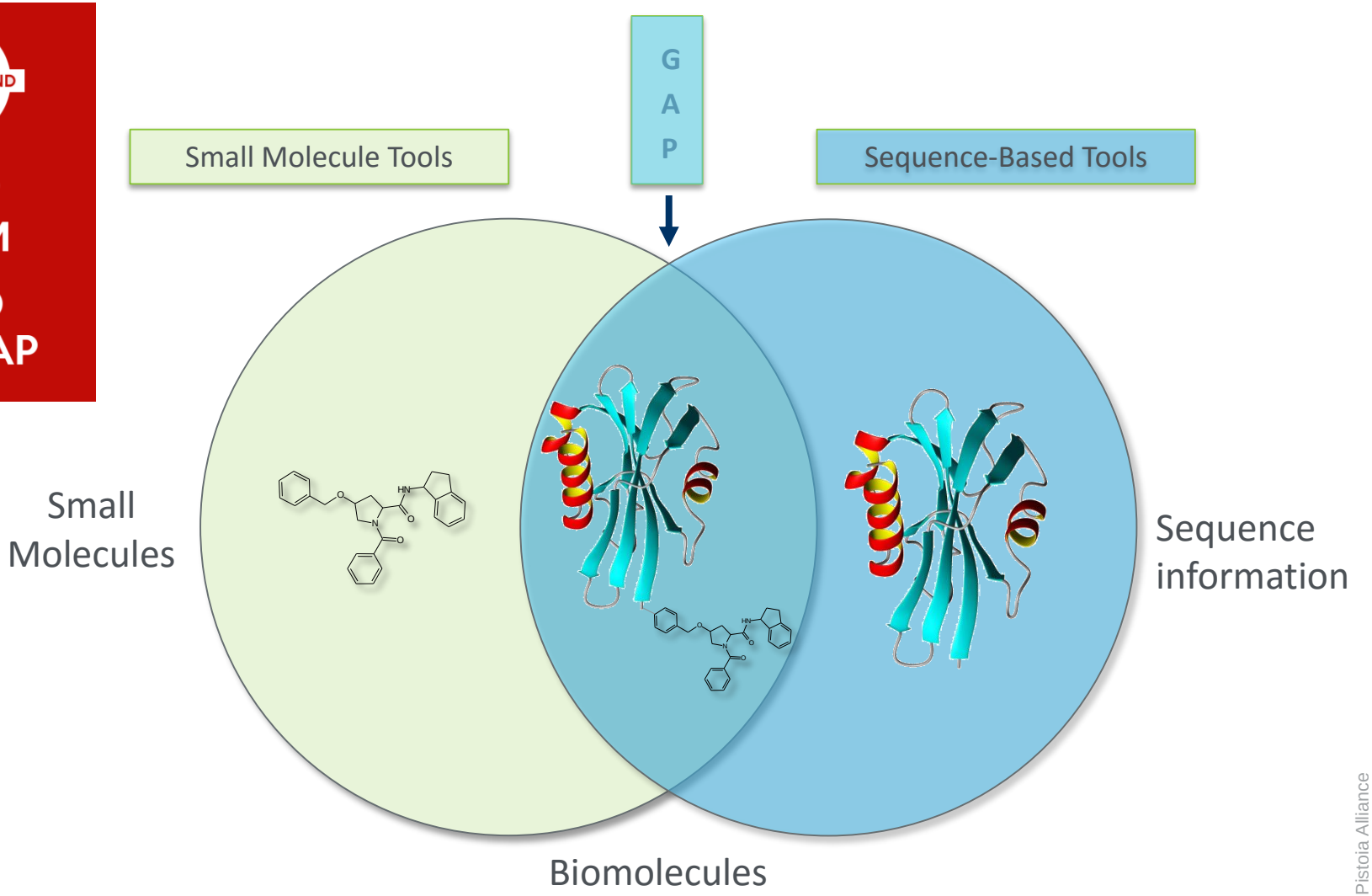
O=C(O)[C@@H](N)C

```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0

  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m1/s1
Key: QNAYBMKLOCPYGJ-UWTATZPHSA-N

Biomolecules - Stuck in the middle…



Mipomersen

An Ionis product used to treat homozygous familial hypercholesterolemia.

The structure can be described as:

G*-C*-C*-U*-C*-dA-dG-dT-dC-dT-dG-dmC-dT-dT-dmC-G*-C*-A*-C*-C*

d = 2'-deoxy
* = 2'-*O*-(2-methoxyethyl)
with phosphorothioate linkages.

This is easy to read

- But relies on a secondary explanation to capture all the information

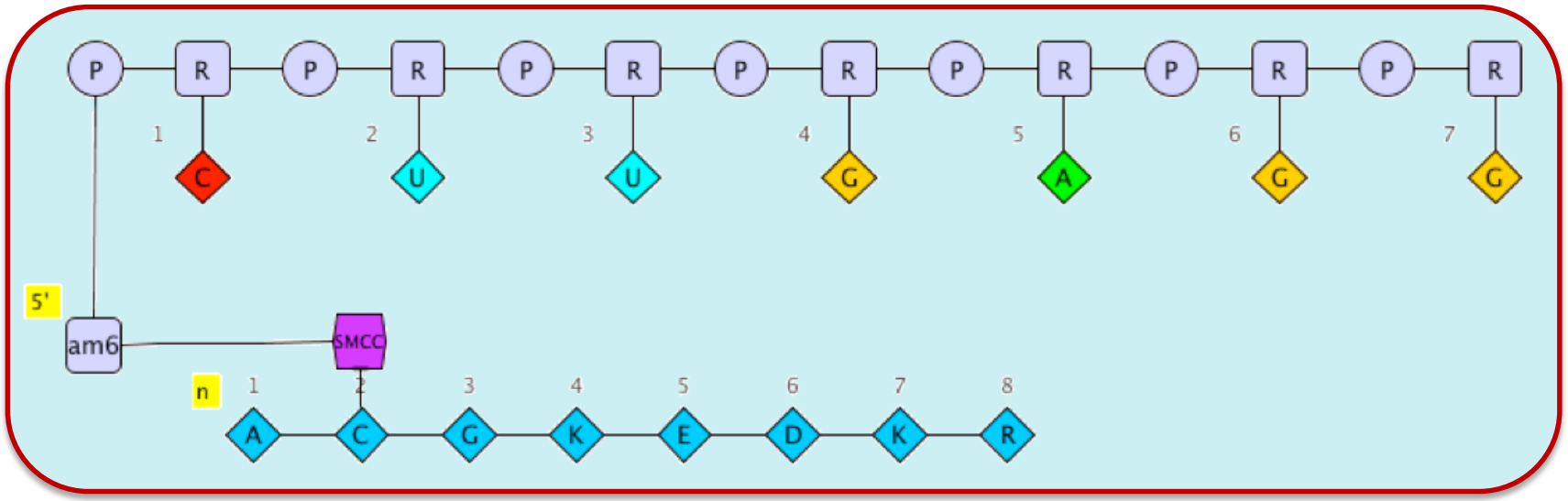
A simple sequence is too limited

Hierarchical **E**diting **L**anguage for **M**acromolecules

Pistoia Alliance

- Hierarchical
  - Biomolecules are “multi-level polymers”

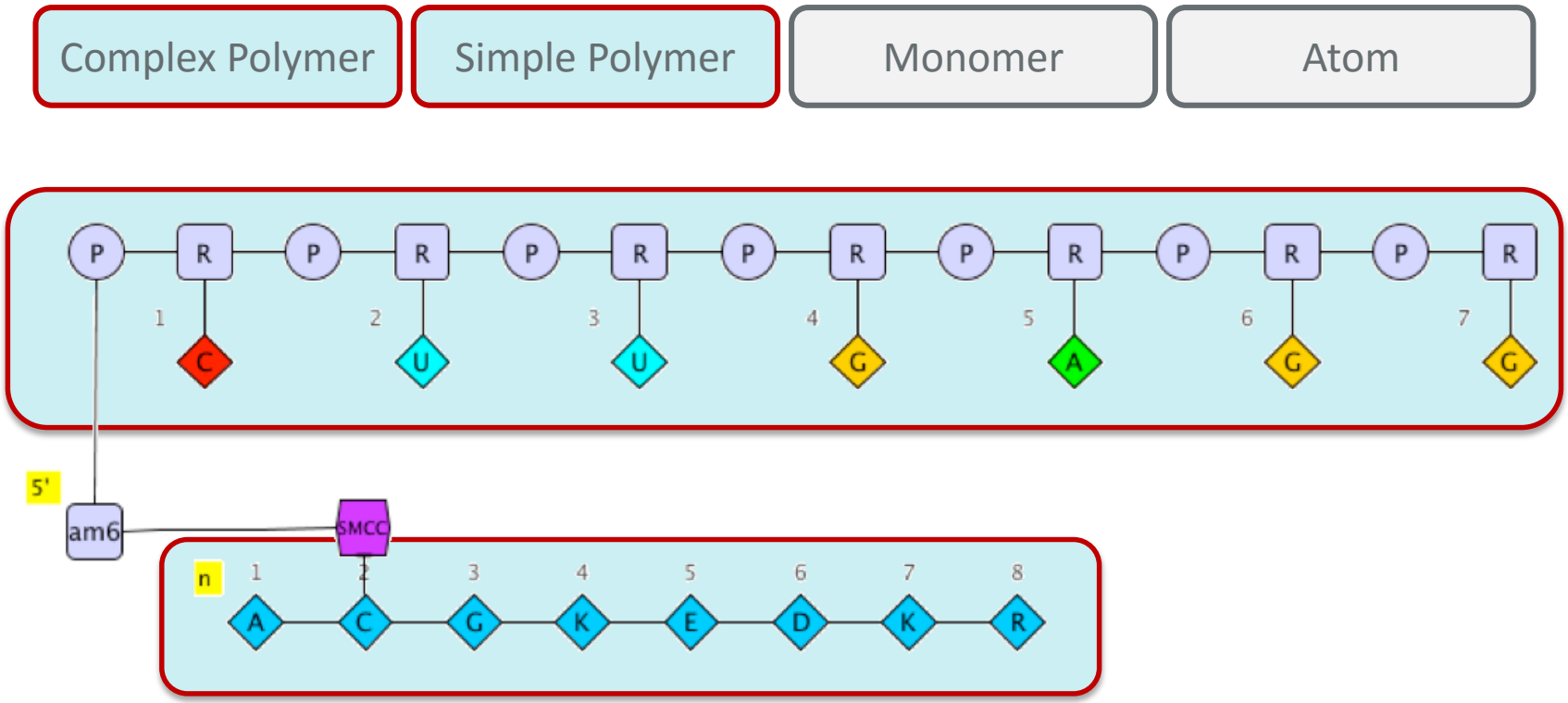
Complex Polymer | Simple Polymer | Monomer | Atom



Hierarchical **E**diting **L**anguage for **M**acromolecules

Pistoia Alliance

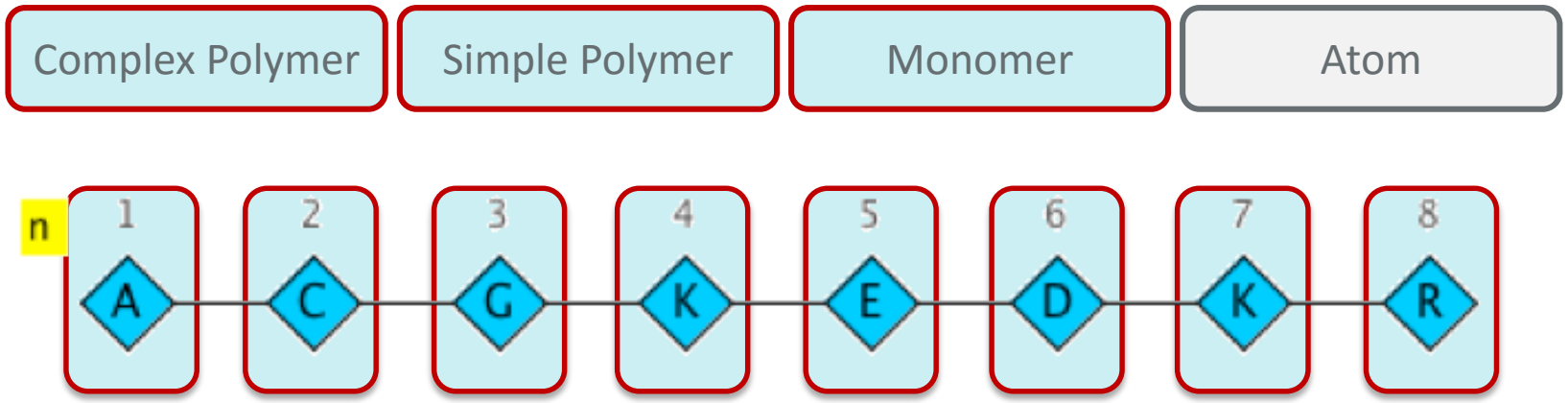
- Hierarchical
  - Biomolecules are “multi-level polymers”



Hierarchical **E**diting **L**anguage for **M**acromolecules

Pistoia Alliance

- Hierarchical
  - Biomolecules are “multi-level polymers”

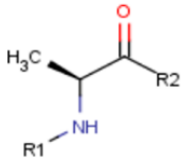


Hierarchical **E**diting **L**anguage for **M**acromolecules

Pistoia Alliance

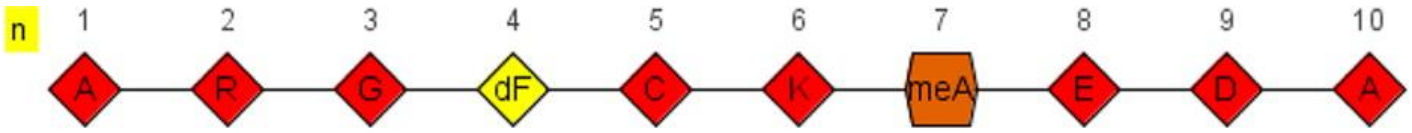
- Hierarchical
  - Biomolecules are “multi-level polymers”

Complex Polymer	Simple Polymer	Monomer	Atom

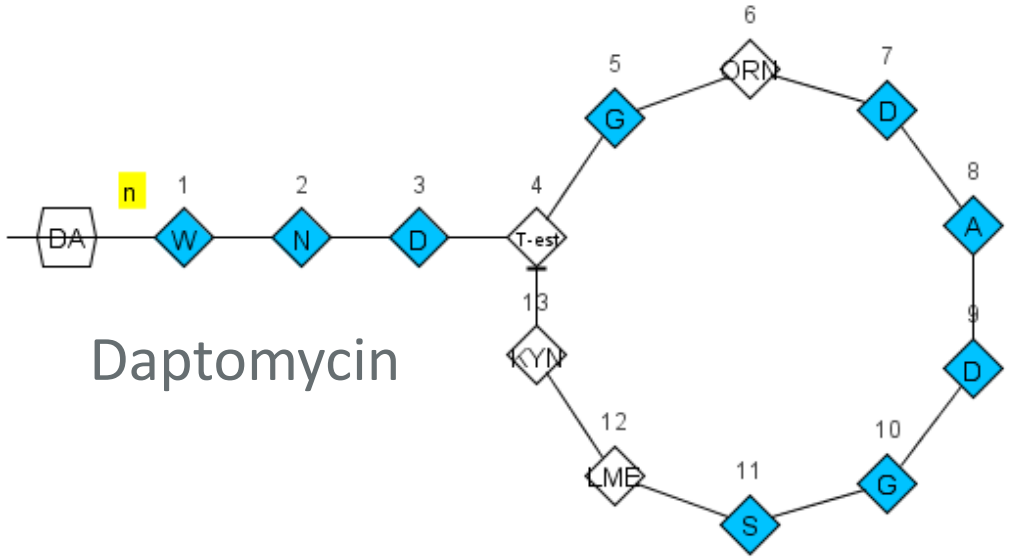
Structure	
SMILES	C[C@H](N[*])C([*])=O |r,$;;;_R1;;_R2;$|
ID	A
Attachment Points	R1-H R2-OH
Natural Analog	A
Polymer Type	PEPTIDE
Monomer Type	Backbone
Name	L-Alanine

Notation format

ListOfSimplePolymers$ListOfConnections$ListOfPolymerGroups$ExtendedAnnotation$



HELM notation: PEPTIDE1{A.R.G.[dF].C.K.[meA].E.D.A}$$$$

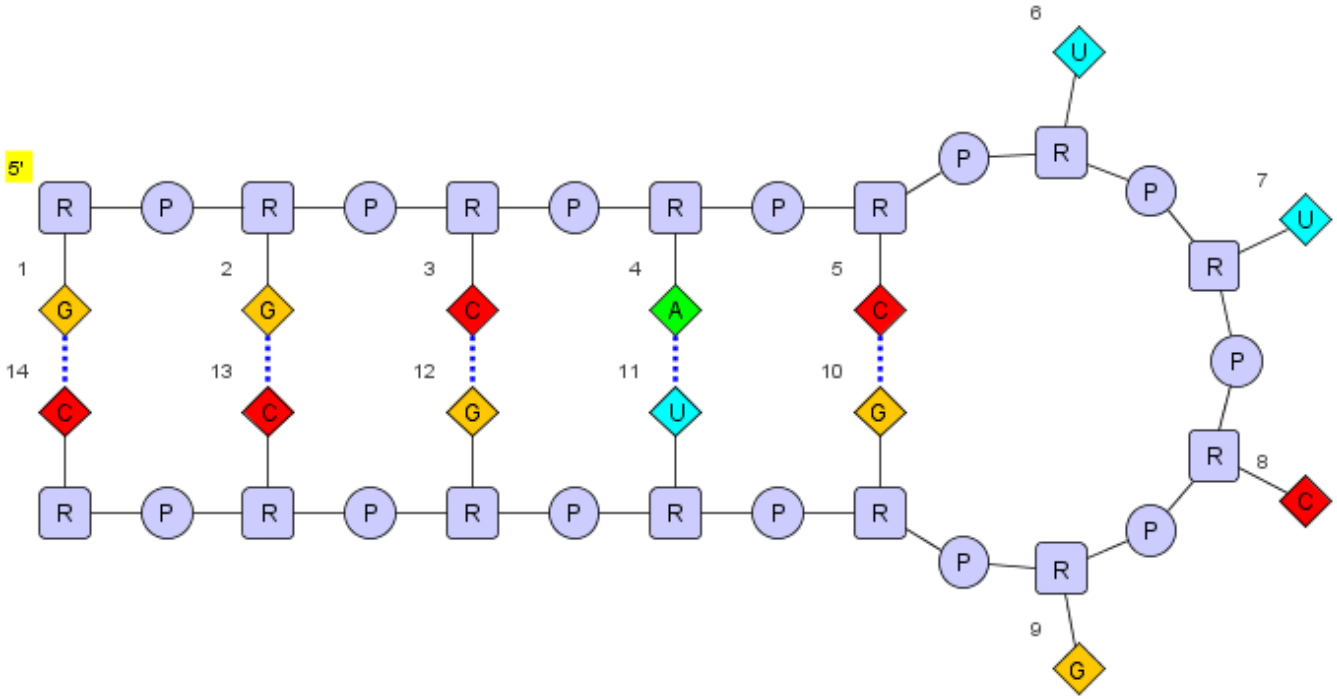


Daptomycin

PEPTIDE1{W.N.D.[T-est].G.[OR].D.A.D.G.S.[LM].[KYN]}|**CHEM1**{DA}$
PEPTIDE1,PEPTIDE1,13:R2-4:R3|PEPTIDE1,CHEM1,1:R1-1:R1$$$

More examples

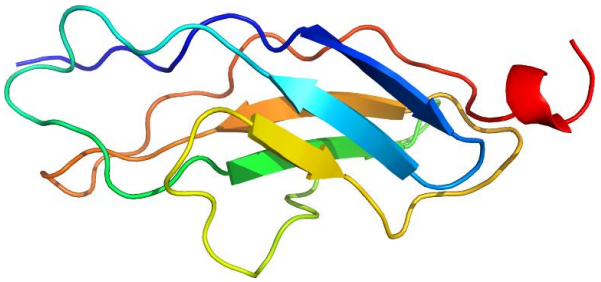
ListOfSimplePolymers$ListOfConnections$ListOfPolymerGroups$ExtendedAnnotation$



HELM notation

RNA1{R(G)P.R(G)P.R(C)P.R(A)P.R(C)P.R(U)P.R(U)P.R(C)P.R(G)P.R(G)P.R(U)P.R(G)P.R(C)P.R(C)}$$RNA1,RNA1,11:pair-32:pair|RNA1,RNA1,5:pair-38:pair|RNA1,RNA1,14:pair-29:pair|RNA1,RNA1,8:pair-35:pair|RNA1,RNA1,2:pair-41:pair$$

HELM…



…and very large molecules

- Example: Connectin (Titin)
- Human muscle protein out of 34,350 amino acids
- => 540,000 atoms
- Creation and validation of HELM notation from FASTA in < 1 s

…and exchanging information

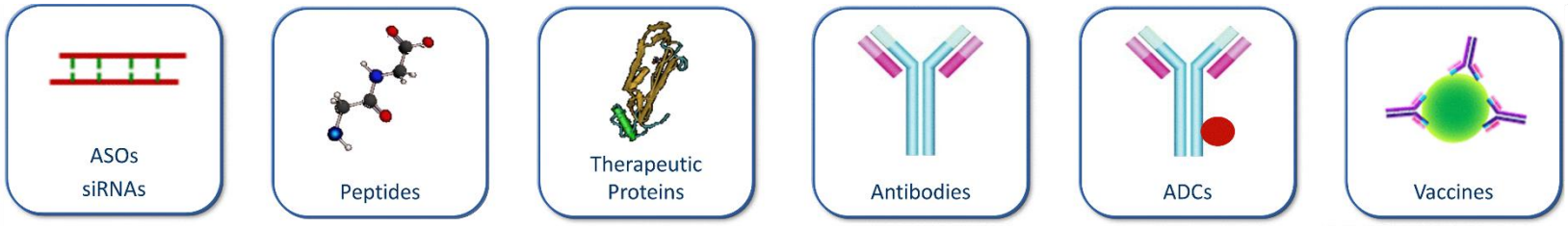
- HELM relies on the monomer definitions
- Monomer information can be exchanged using xHELM, an XML file format that includes the monomers for the molecules in the file.

… and monomer bloat

- In-line HELM allows the definition of ‘temporary’ monomers when you with only use a monomer once.

So… mission completed?

? ? ? ? ? ?

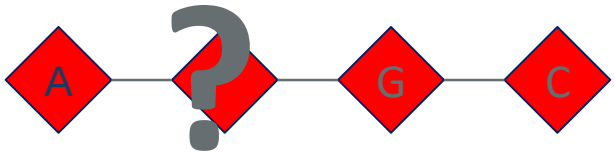


- Unknown numbers of repeating elements
- Various connection points of ADCs
- Unknown elements in sequences
- Unknown connections between polymers
- Undefined polymers
- …

Ambiguity is not something we handle well in either small or large molecule representation.

Monomer Ambiguity

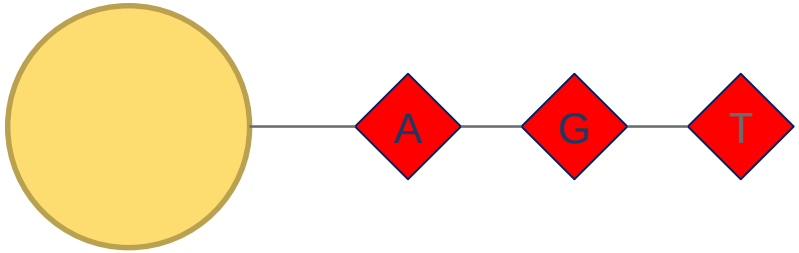
*	0..n unknown monomers	PEPTIDE1{A.*.G.C}$$$$V2.0
X	Single unknown amino acid in a PEPTIDE	PEPTIDE1{A.X.G.C}$$$$V2.0
N	Single unknown base in a RNA	RNA1{R(A)P.R(N)P.R(C)P.R(C)P.R(C)}$$$$V2.0
(,)	One of a list of monomer is possible	PEPTIDE1{A.(A:10,G:90).G.C}$$$$V2.0
(+)	Mixture of monomers	PEPTIDE1{A.(A+G+C).G.C}$$$$V2.0
_	Deleted or missing single monomer	PEPTIDE1{A.(A,_).G.C}$$$$V2.0
´´	Repeating monomers	PEPTIDE1{A.G.A.C.A‘5-30‘}$$$$V2.0



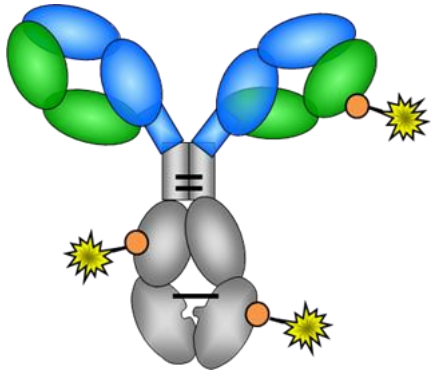
Other Ambiguity Types

Sequence or polymer type is unknown

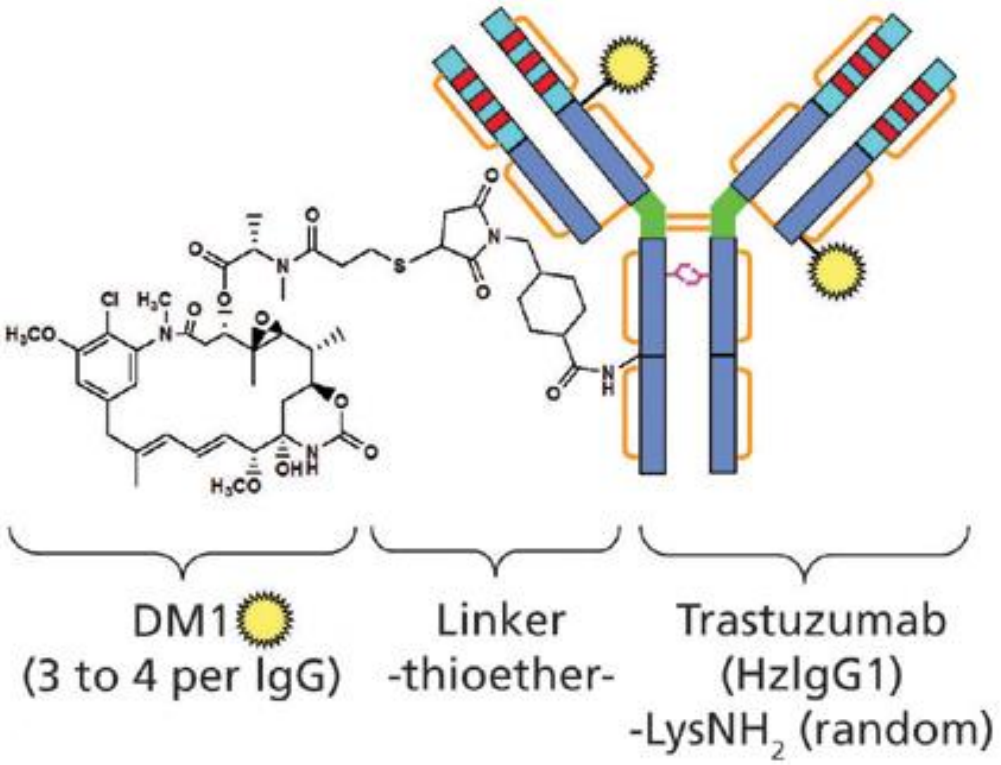
BLOB1{Bead}"Aminated Polystyrene"|PEPTIDE1{A.G.T}$$$$



Connections are unknown



Ambiguity - Kadcyla



- The connection between the drug and antibody is not well understood (Marcoux et al. 2015). Lysine-linked, but where?
- Each trastuzumab molecule may be linked to zero to eight DM1 molecules (3.5 on average)
- Glycosylation is present and not well understood

HELM notation - Kadcyla

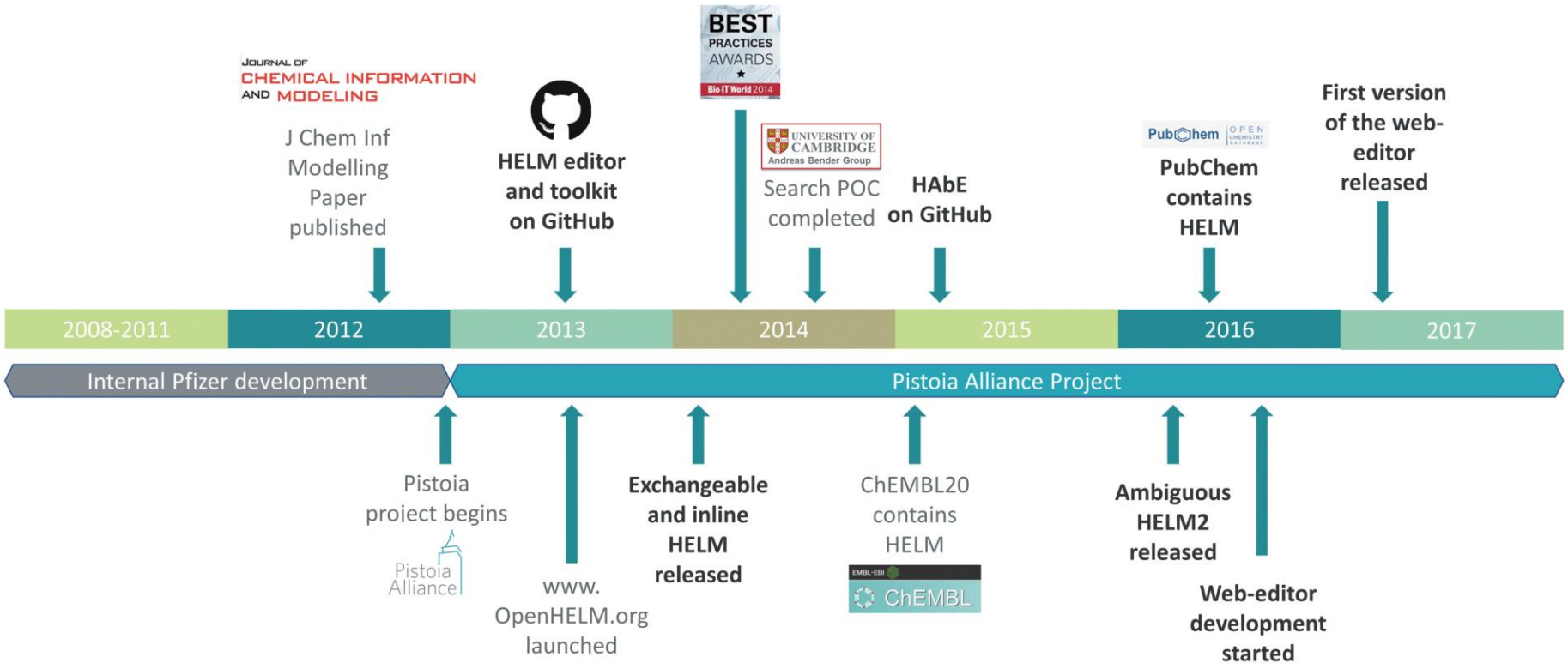
PEPTIDE1{D.I.Q.M.T.Q.S.P.S.S.L.S.A.S.V.G.D.R.V.T.I.T.C.R.A.S.Q.D.V.N.T.A.V.A.W.Y.Q.Q.K.P.G.K.A.P.K.L.L.I.Y.S.A.S.F.L.Y.S.G.V.P.S.R.F.S.G.S.R.S.G.T.D
.F.T.L.T.I.S.S.L.Q.P.E.D.F.A.T.Y.Y.C.Q.Q.H.Y.T.T.P.P.T.F.G.Q.G.T.K.V.E.I.K.R.T.V.A.A.P.S.V.F.I.F.P.P.S.D.E.Q.L.K.S.G.T.A.S.V.V.C.L.L.N.N.F.Y.P.R.E.A
.K.V.Q.W.K.V.D.N.A.L.Q.S.G.N.S.Q.E.S.V.T.E.Q.D.S.K.D.S.T.Y.S.L.S.S.T.L.T.L.S.K.A.D.Y.E.K.H.K.V.Y.A.C.E.V.T.H.Q.G.L.S.S.P.V.T.K.S.F.N.R.G.E.C}|PEPTID
E2{E.V.Q.L.V.E.S.G.G.G.L.V.Q.P.G.G.S.L.R.L.S.C.A.A.S.G.F.N.I.K.D.T.Y.I.H.W.V.R.Q.A.P.G.K.G.L.E.W.V.A.R.I.Y.P.T.N.G.Y.T.R.Y.A.D.S.V.K.G.R.F.T.I.S.A.D
.T.S.K.N.T.A.Y.L.Q.M.N.S.L.R.A.E.D.T.A.V.Y.Y.C.S.R.W.G.G.D.G.F.Y.A.M.D.Y.W.G.Q.G.T.L.V.T.V.S.S.A.S.T.K.G.P.S.V.F.P.L.A.P.S.S.K.S.T.S.G.G.T.A.A.L.G.C
.L.V.K.D.Y.F.P.E.P.V.T.V.S.W.N.S.G.A.L.T.S.G.V.H.T.F.P.A.V.L.Q.S.S.G.L.Y.S.L.S.S.V.V.T.V.P.S.S.S.L.G.T.Q.T.Y.I.C.N.V.N.H.K.P.S.N.T.K.V.D.K.K.V.E.P.P
.K.S.C.D.K.T.H.T.C.P.P.C.P.A.P.E.L.L.G.G.P.S.V.F.L.F.P.P.K.P.K.D.T.L.M.I.S.R.T.P.E.V.T.C.V.V.V.D.V.S.H.E.D.P.E.V.K.F.N.W.Y.V.D.G.V.E.V.H.N.A.K.T.K.P
.R.E.E.Q.Y.N.S.T.Y.R.V.V.S.V.L.T.V.L.H.Q.D.W.L.N.G.K.E.Y.K.C.K.V.S.N.K.A.L.P.A.P.I.E.K.T.I.S.K.A.K.G.Q.P.R.E.P.Q.V.Y.T.L.P.P.S.R.D.E.L.T.K.N.Q.V.S.L
.T.C.L.V.K.G.F.Y.P.S.D.I.A.V.E.W.E.S.N.G.Q.P.E.N.N.Y.K.T.T.P.P.V.L.D.S.D.G.S.F.F.L.Y.S.K.L.T.V.D.K.S.R.W.Q.Q.G.N.V.F.S.C.S.V.M.H.E.A.L.H.N.H.Y.T.Q.K
.S.L.S.L.S.P.G.K}|PEPTIDE3{E.V.Q.L.V.E.S.G.G.G.L.V.Q.P.G.G.S.L.R.L.S.C.A.A.S.G.F.N.I.K.D.T.Y.I.H.W.V.R.Q.A.P.G.K.G.L.E.W.V.A.R.I.Y.P.T.N.G.Y.T.R.Y.A
.D.S.V.K.G.R.F.T.I.S.A.D.T.S.K.N.T.A.Y.L.Q.M.N.S.L.R.A.E.D.T.A.V.Y.Y.C.S.R.W.G.G.D.G.F.Y.A.M.D.Y.W.G.Q.G.T.L.V.T.V.S.S.A.S.T.K.G.P.S.V.F.P.L.A.P.S.S
.K.S.T.S.G.G.T.A.A.L.G.C.L.V.K.D.Y.F.P.E.P.V.T.V.S.W.N.S.G.A.L.T.S.G.V.H.T.F.P.A.V.L.Q.S.S.G.L.Y.S.L.S.S.V.V.T.V.P.S.S.S.L.G.T.Q.T.Y.I.C.N.V.N.H.K.P
.S.N.T.K.V.D.K.K.V.E.P.P.K.S.C.D.K.T.H.T.C.P.P.C.P.A.P.E.L.L.G.G.P.S.V.F.L.F.P.P.K.P.K.D.T.L.M.I.S.R.T.P.E.V.T.C.V.V.V.D.V.S.H.E.D.P.E.V.K.F.N.W.Y.V
.D.G.V.E.V.H.N.A.K.T.K.P.R.E.E.Q.Y.N.S.T.Y.R.V.V.S.V.L.T.V.L.H.Q.D.W.L.N.G.K.E.Y.K.C.K.V.S.N.K.A.L.P.A.P.I.E.K.T.I.S.K.A.K.G.Q.P.R.E.P.Q.V.Y.T.L.P.P
.S.R.D.E.L.T.K.N.Q.V.S.L.T.C.L.V.K.G.F.Y.P.S.D.I.A.V.E.W.E.S.N.G.Q.P.E.N.N.Y.K.T.T.P.P.V.L.D.S.D.G.S.F.F.L.Y.S.K.L.T.V.D.K.S.R.W.Q.Q.G.N.V.F.S.C.S.V
.M.H.E.A.L.H.N.H.Y.T.Q.K.S.L.S.L.S.P.G.K}|PEPTIDE4{D.I.Q.M.T.Q.S.P.S.S.L.S.A.S.V.G.D.R.V.T.I.T.C.R.A.S.Q.D.V.N.T.A.V.A.W.Y.Q.Q.K.P.G.K.A.P.K.L.L.I.Y
.S.A.S.F.L.Y.S.G.V.P.S.R.F.S.G.S.R.S.G.T.D.F.T.L.T.I.S.S.L.Q.P.E.D.F.A.T.Y.Y.C.Q.Q.H.Y.T.T.P.P.T.F.G.Q.G.T.K.V.E.I.K.R.T.V.A.A.P.S.V.F.I.F.P.P.S.D.E
.Q.L.K.S.G.T.A.S.V.V.C.L.L.N.N.F.Y.P.R.E.A.K.V.Q.W.K.V.D.N.A.L.Q.S.G.N.S.Q.E.S.V.T.E.Q.D.S.K.D.S.T.Y.S.L.S.S.T.L.T.L.S.K.A.D.Y.E.K.H.K.V.Y.A.C.E.V.T
.H.Q.G.L.S.S.P.V.T.K.S.F.N.R.G.E.C}|CHEM1{[SMCC]}|CHEM2{[DM1]}|CHEM1,CHEM2,1:R2-1:R1|PEPTIDE1,PEPTIDE1,23:R3-
88:R3|PEPTIDE2,PEPTIDE2,371:R3-429:R3|PEPTIDE4,PEPTIDE4,134:R3-194:R3|PEPTIDE1,PEPTIDE1,134:R3-
194:R3|PEPTIDE3,PEPTIDE3,265:R3-325:R3|PEPTIDE3,PEPTIDE4,224:R3-214:R3|PEPTIDE2,PEPTIDE3,233:R3-
233:R3|PEPTIDE2,PEPTIDE2,265:R3-325:R3|PEPTIDE2,PEPTIDE2,147:R3-203:R3|PEPTIDE2,PEPTIDE3,230:R3-
230:R3|PEPTIDE2,PEPTIDE1,224:R3-214:R3|PEPTIDE3,PEPTIDE3,147:R3-203:R3|PEPTIDE2,PEPTIDE2,22:R3-
96:R3|PEPTIDE3,PEPTIDE3,22:R3-96:R3|PEPTIDE3,PEPTIDE3,371:R3-429:R3|PEPTIDE4,PEPTIDE4,23:R3-88:R3|
G1,CHEM1,K:R3-1:R1$G1(PEPTIDE1+PEPTIDE2+PEPTIDE3+PEPTIDE4)|G2(CHEM1:3.5+G1:1)$$V2.0

Connected to a lysine, but could be any in the group

3.5:1 ratio of SMCC to antibody

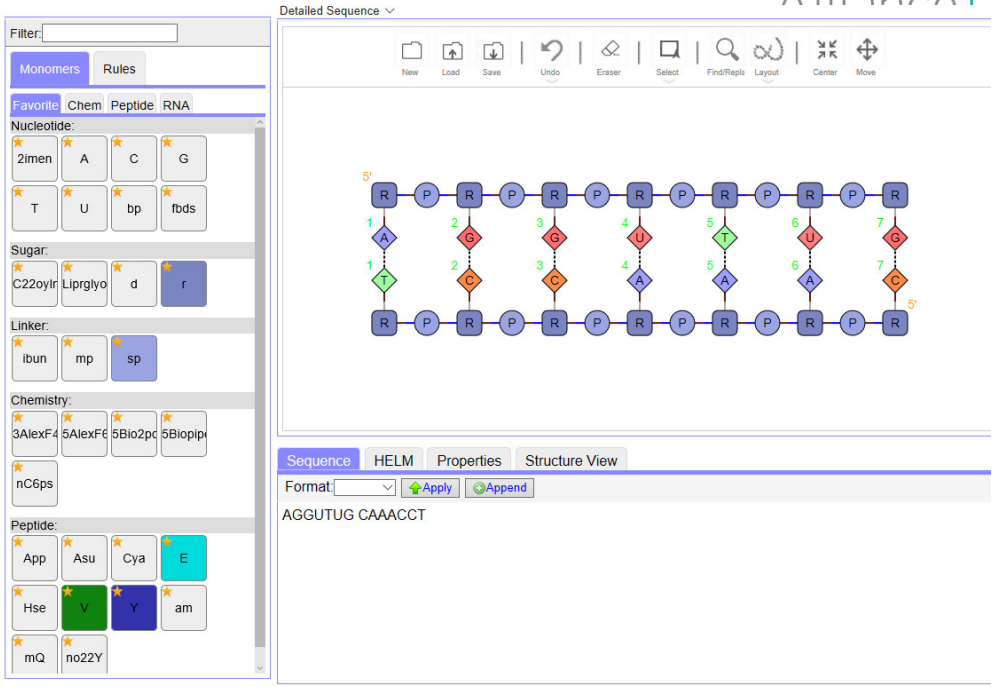
HELM in practice

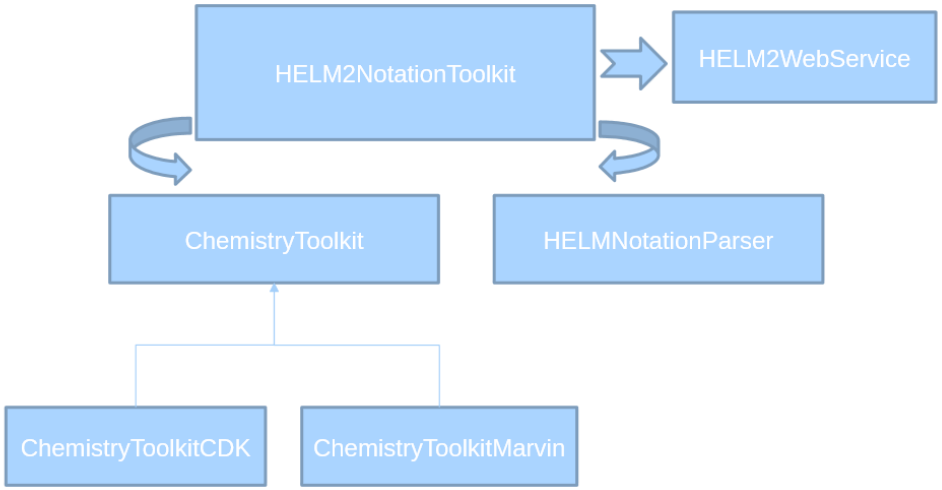
HELM Project



HELM Open Source Tools

Web-based editor – currently HELM1. We are working on HELM2 and expect to publish later this year.

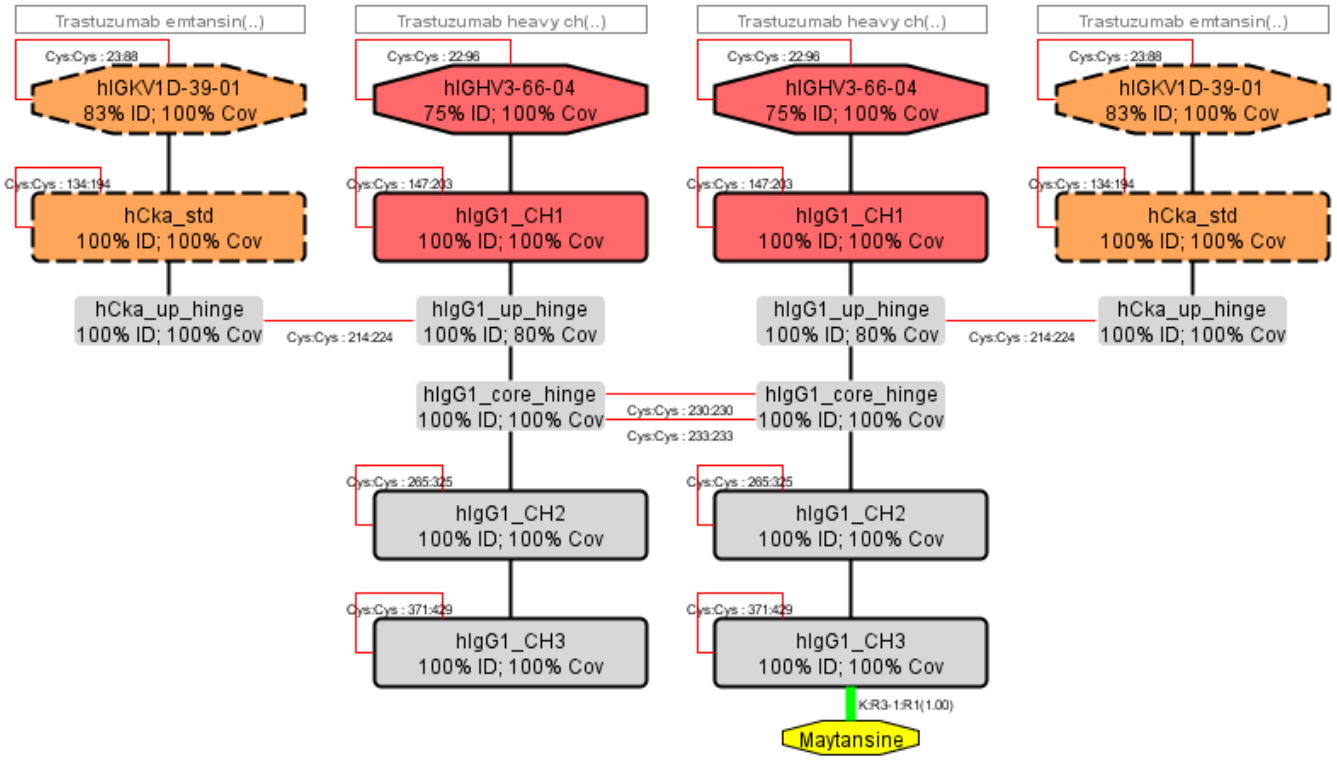




HELM2 is implemented in the toolkit
And can be accessed via RESTFUL web-services

HELM Antibody Editor (HAbE)

HAbE (developed by Roche), will take in an antibody sequence, run a BLAST search and identify and display domains.
The initial domain assignment can be edited if the scientist believes it is incorrect and the editor can be used to adjust Cys-cys bonds and any other structural element.



The HELM Ecosystem

- Pharma / Biotech
  - BMS, GSK, Ionis, Merck, Novartis, Pfizer, Roche
- Software vendors
  - ACD/Labs, BioMax, BIOVIA, ChemAxon, NextMove, PerkinElmer, Scilligence
- Content / Service Providers
  - EBI (ChEMBL), NCBI (PubChem), quattro research
- Regulatory
  - Acceptable format in ISO 11238 guidelines

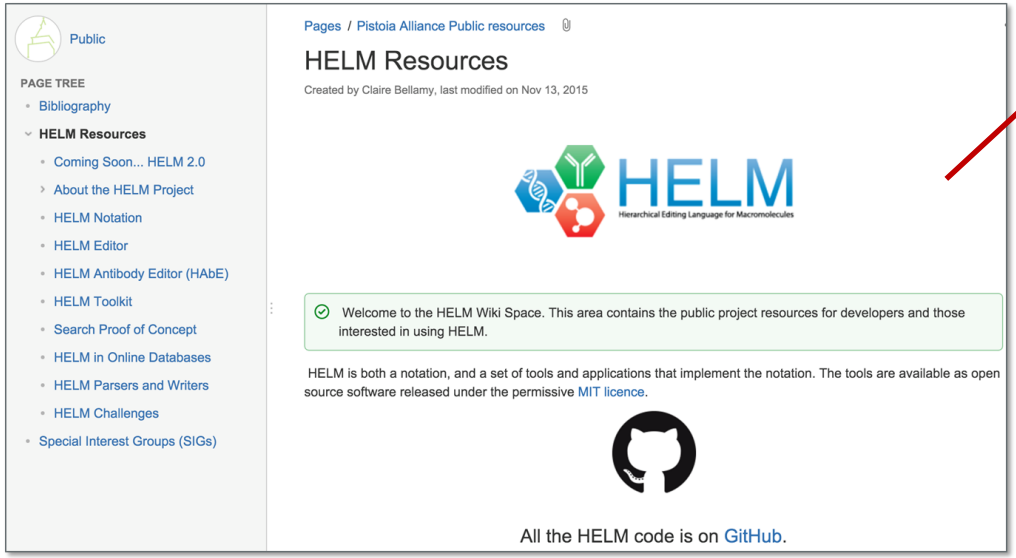
On-line information

www.OpenHelm.org

GitHub

https://github.com/PistoiaHELM

Permissive MIT license



Wiki contains

- Specifications
- User guides
- Presentations
- Links to code

What would Dalton do next?

Monomer naming

Monomers are identified by convention.

- Natural monomers are limited in number and observed in plants and animals.
- Unnatural monomers, are whatever is convenient for synthetic biology.

HELM allows the user to define whatever monomer they like and enables organisations to exchange them, however consistency makes everyone's life easier.

Existing nucleotide conventions

1. IUPAC rules on biochemical nomenclature, Abbreviations and Symbols for Nucleic Acids, Polynucleotides and their Constituents

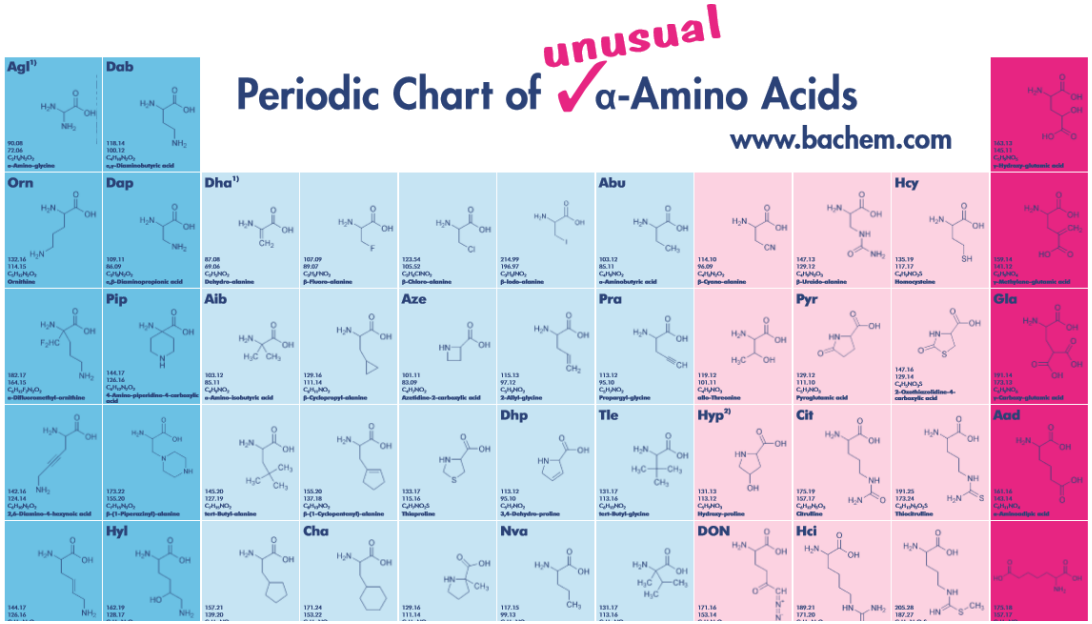
Monomer naming

Peptide monomers

1. IUPAC peptide and amino acid nomenclature
   - Good for D, L amino acids, but does not give much guidance for anything more complex
2. INSDC
   - has a nice table of amino acids and modified ones that use MeGly for N-methylglycine (consortium includes EMBL and GenBank)

Consistent naming conventions are in short supply.

In the absence of anything else – people will make stuff up.



Guidance from serious academic bodies would be helpful.

Ambiguity – a challenge

HELM has made a start, but we are only now building tools to use in ‘real-life’ and may find further areas to work on.

HELM2 can’t be converted into small molecule formats such as molfiles, SMILES and InChI as they can’t handle ambiguity.

InChI are interested in looking at ambiguity and the InChI positional isomers group could be a key enabler

- Shameless plug – this group is looking for new members to help with this work

Acknowledgements

Leadership
- Sergio Rotstein (Pfizer) – Project Lead
- Claire Bellamy (Pistoia Alliance) – Project Manager

Active Team
- Jan Holst Jensen (Chembiofusion)
- Stefan Klostermann (Roche)
- Roland Knispel (ChemAxon)
- Jeff Milton (Ionis)
- Sven Neumeyer (Novartis)
- Matthias Nolte (BMS)
- Yohann Potier (Novartis)
- Joann Precott-Roy (Novartis)
- Eric Swayze (Ionis)
- Markus Weisser (quattro)
- Tianhong Zhang (Pfizer)

Steering Committee
- Margret Assfalg (Roche)
- David Nirschl (BMS)
- Ranjeeva Ramasinghe (GSK)
- Sergio Rotstein (Pfizer)
- Eric Swayze (Ionis)
- John Wise (Pistoia Alliance)
- Quan Yang (Novartis)

Pfizer Team
- Peter Henstock
- David Klatte
- Christine Lawrence
- Frank Loganzo
- Hongli Li
- Sergio Rotstein
- Simone Sciabola
- Rob Stanton
- Nathan Tumey
- Simon Xi
- Tianhong Zhang

A Pistoia Alliance Innovation
HELM
Hierarchical Editing Language for Macromolecules

www.OpenHelm.org
info@openhelm.org

Pistoia Alliance

Thank you

Claire.Bellamy@pistoiaalliance.org

Backup slides