HELM – An emerging standard for the representation of complex biologics

SUMMARY

The increase in sophistication of biotherapeutics research has revealed a gap in informatics methodology whereby small-molecule technologies are too unwieldy for representing large molecules, and sequence-based technologies are incapable of representing the growing variety of biomolecular modalities that include modifications such as unnatural monomers (amino acids, nucleotides) and other chemical modifications (e.g. bioconjugates).

To surmount these challenges internally, researchers at Pfizer developed a hierarchical molecular notation language called HELM that enables them to represent very complex entities including oligonucleotides, peptides, proteins, antibodies and bioconjugates in a flexible and compact fashion. Pfizer has built their biomolecular registration infrastructure, along with many downstream tools, around this technology.

In collaboration with the Pistoia Alliance, HELM was released into the public domain in 2013 and is well on its way to becoming the industry standard for the exchange and manipulation of complex biomolecule structures and their associated data.

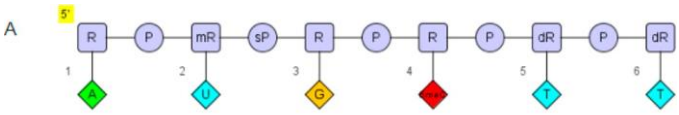
INTRODUCTION

The development of the Hierarchical Editing Language for Macromolecules (HELM) was initiated at Pfizer in 2008 in response to the absence of appropriate technologies for the digital representation of complex therapeutic oligonucleotides.

A paper describing HELM was published in the Journal of Chemical Information and Modeling in September of 2012. Around that same time the Pistoia Alliance voted the problem of biomolecular representation as one of the top challenges faced by its member organizations, thus leading to an initiative to release the HELM technology into the open source community.

RESULTS

HELM uses a hierarchical approach to encode the biomolecule structure. Monomers are defined using a standard atom bond representation such as SMILES, and given an identifier. This identifier is then used to string together sequences of a particular type such as peptide, oligonucleotide, etc. These homogeneous sequences can further be connected to each other to create molecules that contain multiple sequence types.



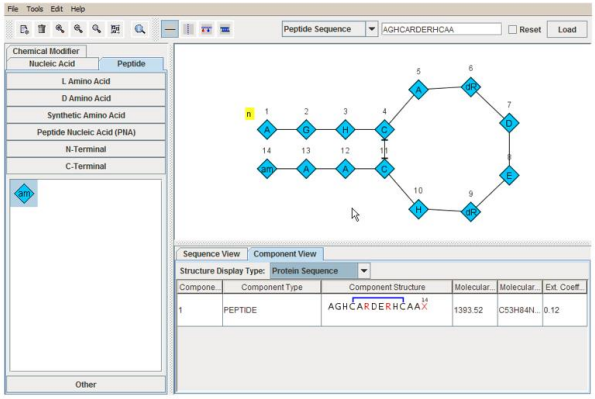
A. Monomer Graph View B. HELM Notation

This hierarchical approach means that the largest, most diverse molecule can be represented in a flexible and compact way. The structure size is reduced by using a monomer IDs in a sequence-like representation but the ability to cope with diversity is preserved by allowing different sequence types to be combined and unique monomers to be defined.

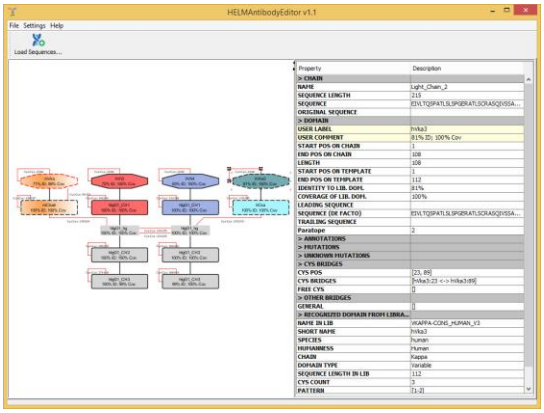
Through the efforts of a uniquely collaborative project team including members from 23 different organizations in the life sciences space, the HELM technology was officially released into the open source community on June 18, 2013. The release consisted of a GitHub repository containing the source code of the toolkit and editor, along with a web site (www.openhelm.org) containing the notation language specification, a free applet version of the editor, training videos and user guides. In 2014 Exchangeable HELM and in-line monomer definitions were implemented and Roche released the HELM Antibody Editor (HAbE) as open source code in March 2015.

The HELM editor consists of a graphical user interface that allows the user to drag and drop monomers onto a canvas, enter the structure as HELM notation or a variety of industry-standard sequence representations. It includes a range of functions supporting common actions such as the

automatic creation of complementary RNA strands. It also includes an editor for the management of monomer definition and structure.



HAbE includes automatic antibody domain recognition and provides a compact domain level view of antibodies and antibody drug conjugates. Cys-Cys bonds can be added automatically or manually and information about the likely species origin is displayed.



ROI

Prior to initiating the HELM project, the Pistoia Alliance commissioned a study which estimated the industry impact of HELM to be an annual productivity saving of over $225m, to be realized through the combination of eliminating the need for other companies to develop their own approach, enabling better information flow between organizations and making more advanced tools available to scientists. This estimated ROI assumes the industry adoption of HELM, something that is still in the early stages, but illustrates the potential of such a game changing innovation.

Uptake has been rapid and includes diverse organizations as content providers, informatics vendors and life science companies. The majority of ChEMBL peptides are now available in HELM, a number of major chemical software vendors have released HELM functionality as part of their suites and two further major pharmaceutical companies have completed internal projects that use HELM to support their internal systems. The project is also aware of a large number of other organizations who are in the process of completing work.

CORE CONTRIBUTING COMPANIES

Biochemfusion, BMS, ChemAxon, GSK, Merck, Novartis, quattro research, Pfizer, Roche.

REFERENCES

HELM website www.openhelm.org

Zhang, Tianhong, et al. "HELM: A hierarchical notation language for complex biomolecule structure representation." *Journal of chemical information and modelling* 52.10 (2012): pp 2796-2806.

Code repository - https://github.com/PistoiaHELM/pistoiahelm.github.com

ABOUT THE PISTOIA ALLIANCE

The Pistoia Alliance is a not-for-profit, cross-company, members' organization, formally incorporated in the State of Delaware. It is committed to lowering the barriers to innovation in life sciences R&D by improving the interoperability of R&D business processes through precompetitive collaboration. It draws its membership widely from life science R&D including Pharma R&D, commercial information providers, technology companies and publicly-funded research institutes. The Pistoia Alliance creates a virtual, open-innovation community by bringing together the key stakeholders to identify the root causes that lead to life science R&D inefficiencies and then develops best practices, technology implementations and standards to overcome these common obstacles.

For further information about the Pistoia Alliance please contact: Claire Bellamy at +44 7508 209365 (UK) and claire.bellamy@pistoiaalliance.org or visit our Web site, http://www.pistoiaalliance.org/