



InChI for large molecules Workshop

Supported by:

IUPAC Division VIII InChI subcommittee

NCBI/NLM

InChI Trust

Lister Hill Center Auditorium
National Library of Medicine
Bldg. 38/Lister Hill Center
1st floor Lobby-Auditorium

Keith T Taylor PhD BSc MRSC
Ladera Consultancy LLC
Sparks, NV

October 27-28 , 2014

Use cases



- Must be quick .. E.g., handle molecules containing up to 15K heavy atoms (1500 residues) in less than one second
- Able to determine the novelty of the chemical entity
- Compare in chemical or sequence based structures
- Can do a search by search engine (e.g., Google)
- Different input formats yields same result (PDB, HELM, SCSR, SMILES, FASTA, MOL/SDF, etc.)
- Can be converted back into output format (PDB, HELM, SCSR, SMILES, FASTA, MOL/SDF, etc.)
- Can handle undefined attachment points of chemical entities (e.g., 1-4 vs. 1-6 in carbohydrates) and variable/undefined stereochemistry (e.g., alpha/beta) and ring open/close variants

Use cases



- Can handle a range of attachments at a defined set of possible locations (e.g., 3 entities with 5 potential places to go)
- Can handle payloads, mutated and modified residues beyond that handled by FASTA
- Be able to group identifiers by sequence
- Handle stereo center variation (L vs. D) for a large number (up to max supported residues)
- Consider arbitrary limit on molecule size (although may have performance implications)
- Be able to retain original sequence information even if chemically modified to be something else (e.g., covalent bonding modification such as cyclization of peptide side chains, etc.)

Use Cases



- Be able to represent complex connectivity with metals, e.g., {cysteine} S-Fe clusters
- Be able to handle peptide/saccharide complexes within a larger complex system, e.g., biological interesting molecules dictionary (BIRD – 1000 cases) .. E.g, be able to handle saccharide cases.
- Handle representation of non-standard polymers found in PTMs, peptides, saccharides, chromophores cases
- Consider generic polymer handling (e.g., undefined overall chemical structure but known components or connection points .. no arbitrary restrictions)
- *

Use cases



- Ensemble molecule with distributions of moieties (e.g., variably described molecule mixture that contains a range of molecular entities that are attached {2-4 of X attached, where X might be a peptide chain})
- Capturing oxidation state of metals complexed with proteins or in nanoparticles
- Must handle well defined large molecules
- Can handle RNA/DNA (nucleic acids) and other biopolymer types that are well defined
- Ability to handle well-defined quat-structure (non-covalently bound, e.g., hemoglobin but not insulin)
- Attempt to preserve stoichiometry of the moieties in question

Use cases



- Ability to ignore hydration from chemistry/sequence description
- Ignore polymorphs (except if stoichiometry is different, do not ignore)
- Consider PEG-ylation aspects (e.g., of proteins and peptides)
- Ability to cover most biopharmaceuticals that are marketed drugs (as-is possible)
- Must be able to handle drugs like defibrotide, heparin
- Handle lipid nanoparticles (e.g., lipidsomes)
- Can handle isotopes (consider cases of variable isotopic enrichment)

High level use cases



- Chemically Modified Biologics exhibit many challenges in chemical representation
 - Size
 - Variable substitution sites
 - Variable substitution loading
 - Hydrogen bonding
 - Presence of heavy metals



Biopolymer testing with InChI v1.05

Keith T Taylor PhD BSc MRSC
Ladera Consultancy LLC
Sparks, NV

Background



- Initial releases of InChI were limited to 1024 heavy atoms
- Many biopolymers of interest contain more than 1024 heavy atoms
- v1.05 removes this limitation and enables InChIs and InChI keys to be calculated for large structures
- This presentation summarizes initial work with large structures using a pre-release version of the software
- The Winchi-1.exe was used to calculate the InChI keys
- Filgrastim sequence was used as the basis for most of the experiments

Limitations



- Structures must be in molfile format
 - V2000 and v3000 formats are accepted
 - v3000 is required for large structures
- The Self Contained Sequence Representation (SCSR) is not supported yet
- Sgroups are not supported and must be removed before presentation to the InChI code
 - Many biopolymer structures contain Sgroup features by default
 - Removal can be achieved programmatically or by editing the molfile in a text editor

Large structure



- Filgrastim

```
      10          20          30          40
 1 M T P L G P A S S L  P Q S F L L K C L E  Q V R K I Q G D G A  A L Q E K L C A T Y
 41 K L C H P E E L V L  L G H S L G I P W A  P L S S C P S Q A L  Q L A G C L S Q L H
 81 S G L F L Y Q G L L  Q A L E G I S P E L  G P T L D T L Q L D  V A D F A T T I W Q
121 Q M E E L G M A P A  L Q P T Q G A M P A  F A S A F Q R R A G  G V L V A S H L Q S
161 F L E V S Y R V L R  H L A Q P
```

- InChIKey=KOKXRWZWQJXBOP-NJDFSSKJBA-N

With disulfide bridges

- InChIKey=MMCZGSMNPYTOPN-NJDFSSKJBA-N

10	20	30	40
MTPLGPASSL	PQSFLLK C LE	QVRKIQGDGA	ALQEKL C ATY
41 KLC H PEELVL	LGHSLGIPWA	PLSSCPSQAL	QLAGC L SQLH
81 SGLFLYQGLL	QALEGISPEL	GPTLDTLQLD	VADFATTIWQ
121 QMEEELGMAPA	LQPTQGAMPA	FASAFQRRAAG	GVLVASHLQS
161 FLEVSYRVLR	HLAQHP		

- InChIKey=MEMBSBQMAVSGHQ-NJDFSSKJBA-N

10	20	30	40
MTPLGPASSL	PQSFLLK C LE	QVRKIQGDGA	ALQEKL C ATY
41 KLC H PEELVL	L GHSLGIPWA	PLSSCPSQAL	QLAGC L SQLH
81 SGLFLYQGLL	QALEGISPEL	GPTLDTLQLD	VADFATTIWQ
121 QMEEELGMAPA	LQPTQGAMPA	FASAFQRRAAG	GVLVASHLQS
161 FLEVSYRVLR	HLAQHP		

Cyclized

- InChIKey=IZNXXFOUFDSLAX-VBNFVGOYBA-N

1	M	T P L G P A S S L	P Q S F L L K C L E	Q V R K I Q G D G A	A L Q E K L C A T Y	10 20 30 40
41	K L C H P E E L V L	L G H S L G I P W A	P L S S C P S Q A L	Q L A G C L S Q L H		
81	S G L F L Y Q G L L	Q A L E G I S P E L	G P T L D T L Q L D	V A D F A T T I W Q		
121	Q M E E L G M A P A	L Q P T Q G A M P A	F A S A F Q R R A G	G V L V A S H L Q S		
161	F L E V S Y R V L R	H L A Q P				

- InChIKey=IZNXXFOUFDSLAX-VBNFVGOYBA-N

1	P	L S S C P S Q A L	Q L A G C L S Q L H	S G L F L Y Q G L L	Q A L E G I S P E L	10 20 30 40
41	G P T L D T L Q L D	V A D F A T T I W Q	Q M E E L G M A P A	L Q P T Q G A M P A		
81	F A S A F Q R R A G	G V L V A S H L Q S	F L E V S Y R V L R	H L A Q P M T P L G		
121	P A S S L P Q S F L	L K C L E Q V R K I	Q G D G A A L Q E K	L C A T Y K L C H P		
161	E E L V L L G H S L	G I P W A				

Multiple cyclizations

- InChIKey=AQUGLJGKXYTOSD-VBNFVGOYBA-N

1	P L S S C P S Q A L Q L A G C L S Q L H	10	20	30	40
41	G P T L D T L Q L D V A D F A T T I W Q Q M E E L G M A P A L Q P T Q G A M P A				
81	F A S A F Q R R A G G V L V A S H L Q S F L E V S Y R V L R H L A Q P M T P L G				
121	P A S S L P Q S F L L K C L E Q V R K I Q G D G A A L Q E K L C A T Y K L C H P				
161	E E L V L L G H S L G I P W A				

The sequence is shown in a table with 5 rows. The first row contains positions 1 through 40. The second row contains positions 41 through 80. The third row contains positions 81 through 120. The fourth row contains positions 121 through 160. The fifth row contains positions 161 through 161. Red brackets highlight specific segments: one from position 1 to 10, another from 10 to 30, and a third from 30 to 40. A blue bracket highlights a segment from position 121 to 130.

Reversed sequence



- InChIKey=YFXNVYXMKDIHRN-VBNFVGOYBA-N

10 20 30 40
OH1 M T P L G P A S S L P Q S F L L K C L E Q V R K I Q G D G A A L Q E K L C A T Y
41 K L C H P E E L V L L G H S L G I P W A P L S S C P S Q A L Q L A G C L S Q L H
81 S G L F L Y Q G L L Q A L E G I S P E L G P T L D T L Q L D V A D F A T T I W Q
121 Q M E E L G M A P A L Q P T Q G A M P A F A S A F Q R R A G G V L V A S H L Q S
161 F L E V S Y R V L R H L A Q P **H**

Filgrastim Lys10-D form



- InChIKey=KOKXRWZWQJXBOP-FNWNWACTBA-N

1	M T P L G P A S S 1	10	P Q S F L L K C L E	20	Q V R K I Q G D G A	30	A L Q E K L C A T Y	40
41	K L C H P E E L V L	L G H S L G I P W A	P L S S C P S Q A L	Q L A G C L S Q L H				
81	S G L F L Y Q G L L	Q A L E G I S P E L	G P T L D T L Q L D	V A D F A T T I W Q				
121	Q M E E L G M A P A	L Q P T Q G A M P A	F A S A F Q R R A G	G V L V A S H L Q S				
161	F L E V S Y R V L R	H L A Q P						

Synthetic Erythropoietin

- InChIKey=XJBDLLBKUYAKW-WAXLMBMOBA-D

1	A P P R L I C D S R	V L E R Y L L E A K	E A K I T T G C A E	H C S L N E K I T V	10	20	30	40
41	P D T K V N F Y A W	K R M E V G Q Q A V	E V W Q G L A L L S	E A V L R G Q A L L				
81	V K S S Q P W C P L	Q L H V D K A V S G	L R S L T T L L R A	L G A Q K C A I S P				
121	P D A A K A A P L R	T I T A D T F R K L	F R V Y S N F L R G	K L K L Y T G E A C				
161	R T G D R							

- PEGylated at K23 and K125
- Acylated at C88 and C106

Polynucleotide - 1



- InChIKey=NELTZQNSFHRPGO-AZBJDUHQBA-N

1	CGGAGCCTGC	10	AGCCCCAGCCC	20	CACCCAGACC	30	CATGGCTGGA	40	CCTGCCACCC
51	AGAGCCCCAT		GAAGCTGATG		GCCCTGCAGC		TGCTGCTGTG		GCACAGTGCA
101	CTCTGGACAG		TGCAGGAAGC		CACCCCCCTG		GGCCCTGCCA		GCTCCCTGCC
151	CCAGAGCTTC		CTGCTCAAGT		GCTTAGAGCA		AGTGAGGAAG		ATCCAGGGCG
201	ATGGCGCAGC		GCTCCAGGAG		AAGCTGGTGA		GTGAGTGTC		CACCTACAAG
251	CTGTGCCACC		CCGAGGGAGCT		GGTGCTGCTC		GGACACTCTC		TGGGCATCCC

- Calculation time: ~6s
- Molecular Formula: C₂₈₉₄H₃₆₄₉N₁₁₄₇O₁₇₉₁P₃₀₀

Polynucleotide - 2



- InChIKey=IHDBOWIRPNDUCX-YUQJSOSJBA-N

1	CGGAGCCTGC	10	AGCCCAGCCC	20	CACCCAGACC	30	CATGGCTGGA	40	CCTGCCACCC	50
51	AGAGCCCCAT		GAAGCTGATG		GCCCTGCAGC		TGCTGCTGTG		GCACAGTGCA	
101	CTCTGGACAG		TGCAGGAAGC		CACCCCCCTG		GGCCCTGCCA		GCTCCCTGCC	
151	CCAGAGCTTC		CTGCTCAAGT		GCTTAGAGCA		AGTGAGGAAG		ATCCAGGGCG	
201	ATGGCGCAGC		GCTCCAGGAG		AAGCTGGTGA		GTGAGTGTGC		CACCTACAAG	
251	CTGTGCCACC		CCGAGGAGCT		GGTGCTGCTC		GGACACTCTC		TGGGCATCCC	
301	CTGGGCTCCC		CTGAGCAGCT		GCCCCAGCCA		GGCCCTGCAG		CTGGCAGGCT	
351	GCTTGAGCCA		ACTCCATAGC		GGCCTTTTCC		TCTACCAGGG		GCTCCCTGCAG	
401	GCCCTGGAAG		GGATCTCCCC		CGAGTTGGGT		CCCACCTTGG		ACACACTGCA	
451	GCTGGACGTC		GCCGACTTTG		CCACCAACCAT		CTGGCAGCAG		ATGGAAGAAC	
501	TGGGAATGGC		CCCTGCCCTG		CAGCCCACCC		AGGGTGCCAT		GCCGGCCTTC	
551	GCCTCTGCTT		TCCAGCGCCG		GGCAGGAGGG		GTCCTGGTTG		CCTCCCCATCT	

- Calculation time: ~38s
- Molecular Formula: C₅₇₈₂H₇₃₀₅N₂₂₅₅O₃₆₀₂P₆₀₀

Polynucleotide - 3



- InChIKey=

1	CGGAGCCCTGC	10	AGCCCAGCCC	20	CACCCAGACC	30	CATGGCTGGA	40	CCTGCCACCC	50
51	AGAGCCCCAT		GAAGCTGATG		GCCCTGCAGC		TGCTGCTGTG		GCACAGTGCA	
101	CTCTGGACAG		TGCAGGAAGC		CACCCCCCTG		GGCCCTGCCA		GCTCCCTGCC	
151	CCAGAGCTTC		CTGCTCAAGT		GCTTAGAGCA		AGTGAGGAAG		ATCCAGGGCG	
201	ATGGCGCAGC		GCTCCAGGAG		AAGCTGGTGA		GTGAGTGTGC		CACCTACAAG	
251	CTGTGCCACC		CCGAGGGAGCT		GGTGCTGCTC		GGACACTCTC		TGGGCATCCC	
301	CTGGGCTCCC		CTGAGCAGCT		GCCCCAGCCA		GGCCCTGCAG		CTGGCAGGCT	
351	GCTTGAGCCA		ACTCCATAGC		GGCCTTTTC		TCTACCAGGG		GCTCCTGCAG	
401	GCCCTTGAAG		GGATCTCCCC		CGAGTTGGGT		CCCACCTTGG		ACACACTGCA	
451	GCTGGACGTC		GCGGACTTTG		CCACCCACCAT		CTGGCAGCAG		ATGGAAGAAC	
501	TGGAATGGC		CCCTGCCCTG		CAGCCCACCC		AGGGTGCCAT		GCCGGCCTTC	
551	GCCTCTGTT		TCCAGCGCCG		GGCAGGAGGG		GTCCTGGTTG		CCTCCCCATCT	
601	GCAGAGCTTC		CTGGAGGTGT		CGTACCGCGT		TCTACGCCAC		CTTGGCCAGC	
651	CCTGAGCCAA		GCCCTCCCCA		TCCCATGTAT		TTATCTCTAT		TTAATATTAA	
701	TGTCTATTAA		AGCCTCATAT		TTAAAGACAG		GGAAAGAGCAG		AACGGAGCCC	
751	CAGGCCTCTG		TGTCCCTCCC		TGCATTTCTG		AGTTTCATTC		TCCTGCCTGT	
801	AGCAGTGAGA		AAAAGCTCCT		GTCCTCCCCAT		CCCCCTGGACT		GGGAGGTAGA	
851	TAGGTAAATA		CCAAGTATTT		ATTACTATGA		CTGCTCCCCA		GCCCTGGCTC	

- Calculation timeout at ~125s
- Molecular Formula: C₈₆₉₃H₁₀₉₈₉N₃₃₂₅O₅₄₂₀P₉₀₀

Myosin-1



- InChIKey=BBJMARUZQDWUQG-PZLOAVSTBA-N

- Calculation time: ~94s
- Molecular Formula: C₉₇₂₅H₁₅₈₁₆N₂₇₄₈O₃₁₀₀S₇₂

1 MSSDSEMAIF 10 GEAAPFLRKS 20 ERERIEAQNK 30 PFDAKTSVVF 40 VDPKESFVK
51 TVQSREGGKV 50 TAKTEAGATV TVKDDQVFPM NPPKYDKIED MAMMTHLHEP
101 AVLYNLKERY AAWMIYTYSG 151 LFCVTVNPYK WLPVYNAEVV TAYRGKKRQE
151 APPHIFSISID 151 NAYQFMLTDR ENQSILITGE SGACKVNTK RVIQYFATIA
201 VTGEKKKEEV 201 TSGKMQGTLE DQIISANPLL EAEGNAKTVR NDNSSRFGKF
251 IRIHFGTTGK 251 LASADIETYL LEKSRTVFQL KAERSYHIFY QIMSNKKPDL
301 IEMLLITTNP 301 YDYAFVSQGE ITVPSIDDQE ELMATDSAIE ILGFTSDERV
351 SIYKLTGAVM 351 HYGNMKFKQK QREEQAEPDG TEVADKAAYL QNLSADLLK
401 ALCYPRVKVG 401 NEYVTKGQTV QQVYNAVAGAL AKAVYDKMFL WMVTRINQQL
451 DTKQPRQYFI 451 GVLDIAGFEI FDFNSLEQLC INFTNEKLQQ FFNHHMFVLE
501 QEEYKKEGIE 501 WTFIDFGMDL AACIELIEKP MGIFSILEE CMEPKATDTS
551 FKNKLYEQHL 551 GKSNNFQKPK PAKGKREAHF SLIHAGTVD YNIAGWLDKN
601 KDPLNETVVG 601 LYQKSAMKTL ALLFVGATGA EAEAGGGKG GKKGSSSFQT
651 VSALFRENLN 651 KLMTNLRSTH PHFVRCIIPN ETKTPGAMEH ELVLHQLRCN
701 GYLEGIRICR 701 KGFPSPRILYA DFQKRYKVLN ASAIPEGQFI DSKKASEKLL
751 GSIDIDHTQY 751 KFGHTKVFFK AGLLGLLEEM RDEKLAQLIT RTQAMCRGFL
801 ARVEYQKMVE 801 RRESIFCIQY NVRAFMNVKH WPWMKLYFKI KPLLKSAETE
851 KEMANMKEEF 851 EKTKEELAKT EAKRKELEEK MVTLMQEKND LQLQVQAEAD
901 SLADAERCD 901 QLIKTKIQLE AKIKEVTERA EDEEEINAEL TAKKRKLEDE
951 CSELKKDIDD 951 LELTLAKVEK EKHATENVKV NLTEEMAGLD ETIAKLTKEK
1001 KALQEAHQQT 1001 LDDLQAEEDK VNLTAKAKIK LEQQVDDLEG SLEQEKKIRM
1051 DLERAKRKLE 1051 GDLKLAQEST MDIENDKQQL DEKLKKKEFE MSGLQSKIED
1101 EQALGMQLQK 1101 KIKELQARIE ELEEEEIAER ASRAKAEKQR SDLSRELEEI
1151 SERLEEAGGA 1151 TSAQIEMNKK REAEFQKMRRT DLEEATLQHE ATAATLRKKH
1201 ADSVAELGEQ 1201 IDNLQRVKQK LEKEKSEMKM EIDDLASNME TVSKAKGNLE
1251 KMCRALEDQL 1251 SEIKTKEEEEQ QRLINDLTAQ RARLQTESGE YSRQLDEKDT
1301 LVSQLSRGKQ 1301 AFTQQIEELK RQLEEEIKAK SALAHALQSS RHDCDLLREQ
1351 YEEEQEAKE 1351 LQRAMSKANS EVAQWRTKYE TDAIQRTEEL EEAKKKLAQR
1401 LQDAEEHVEA 1401 VNAKCASLEK TKQRLQNEVE DLMIDVERTN AACAAALDKKQ
1451 RNFDKILAWE 1451 KQKCEETHAE LEASQKESRS LSTELFKIKN AYEESLDQLE
1501 TLKRENKNLQ 1501 QEISDLTEQI AEGGKRIHEL EKIKKQVEQE KSELQAALEE
1551 AEASLEHEEG 1551 KILRIQLELN QVKSEVDRKI AEKDEEIDQM KRНHIRIVES
1601 MQSTLDAEIR 1601 SRNDAIRLKK KMEGDLNEME IQLNHNARMA AEALRNYRNT
1651 QAILKDTQLH 1651 LDDALRSQED LKEQLAMVER RANLLQAEIE ELRATLQE
1701 RSRKIAEQEL 1701 LDASERVQLL HTQNTSLINT KKLETDISQ IQGEMEDIQ
1751 EARNAEAKK 1751 KAITDAAMMA EELKKEQDTS AHLERMKKNL EQTVKDLQHR
1801 LDEAEQLALK 1801 GGKKQIQLKE ARVRELEGEV ESEQKRNVEA VKGLRKHERK
1851 VKELTYQTEE 1851 DRKNILRLQD LVDKLQAKVK SYKRQAEAE EQSNVNLSKF
1901 RRIQHELEEEA 1901 EERADIAESQ VNKLRLVKSRE VHTKIISEE

Trastuzumab dimer

- InChIKey=VRBUFPXQWJVPL0-JNJMYDJTBA-N
- Single arbitrary stereocenter inverted
 - InChIKey=VRBUFPXQWJVPL0-RCEINSQCBA-N
- Calculation time: ~27s
- Molecular Formula: C₆₄₆₀H₉₉₇₂N₁₇₂₄O₂₀₁₄S₄₄

<img alt="Diagram showing two identical protein sequences for Trastuzumab dimer, each with 201 amino acids. The sequences are aligned vertically. Red boxes highlight specific residues at positions 10, 20, 30, 40, and 50. The first sequence has a red bracket under the first 20 residues, and the second sequence has a red bracket under the last 20 residues. The sequences are:

Sequence 1 (Top):
1 EVQLVESGGG LVQPGGSLRL SCAASGFNIK DTYIHWRQA PGKGLEVAR
51 IYPTNGYTRY ADSVKGRFTI SADTSKNTAY LQMNSLRAED TAVYYC SRWG
101 GDGFYAMDYW GOGTLVTVSS ASTKGPSVFP LAPSSKSTSG GTAALGC LVK
151 DYFPEPVTVS WNSGALTSGV HTFPAVLQSS GLYSLSSVVT VPSSSLGTQT
201 YICCNVNHKPS NTKVDKKVEP KSCDKTHTC P PCPAPELLGG PSVFLFPPKP
251 KDTLMISRTP EVTCCVVVDVS HEDPEVKFNW YVDGVEVHNA KTKPREEQYN
301 STYRVVSVLT VLHQDWLNGK EYKC KVSNKA LPAPIEKTI S KAKGQPREPQ
351 VYTLPPSREE MTKNQVSLTC LVKGFYPSDI AVEWESENQGP ENNYKTTPPV
401 LDSDGSFFLY SKLTVDKSRW QQGNVFSC SV MHEALHNHYT QKSLSLSPGK

Sequence 2 (Bottom):
1 DIQMTQSPSS LSASVGDRVT ITCRASQDVN TAVAWYQQKP GKAPKLLIYS
51 ASFLYSGVPS RFSGSRSGTD FTLTISSSLQF EDFATIYYC QQ HYTPPPTFGQ
101 GTKVEIKRTV AAPSVFIFPP SDEQLKSGTA SVVC LLNNFY PREAKVQWKV
151 DNALQSGNSQ ESVTEQDSKD STYSLSSTLT LSKADYEKHK VYAC EVTHQG
201 LSSPVTKSFN RGE C

The diagram shows a vertical alignment of the two sequences. Red boxes highlight specific residues at positions 10, 20, 30, 40, and 50. The first sequence has a red bracket under the first 20 residues, and the second sequence has a red bracket under the last 20 residues.

Summary



- InChI v1.05 can generate InChI strings and keys from large structures
 - InChI strings are unwieldy
- All calculations were done using the winchi-1 application
 - Convenient to use but not the most efficient method for calculating InChI strings and keys
- Calculation time for Filgrastim related peptides and the synthetic erythropoietin were not perceptible using the winchi-1 program
- Processing time for large structures needs to be improved
- A large polynucleotide timed out but a polypeptide of similar size did not
- Myosin-1, presented as a linear peptide, took ~94s to process whereas Trastuzumab took ~27s
- Canonicalization may be an area of weakness
- Trastuzumab stereoisomers are differentiated
 - An arbitrary stereocenter in Trastuzumab was inverted
 - Processing time unchanged
 - Different InChI key was produced



Next Steps

Keith T Taylor PhD BSc MRSC
Ladera Consultancy LLC
Sparks, NV

Trastuzumab emtansine: patent extract



The mertansine is conjugated to the trastuzumab through a maleimidocaproyl (MC) linker which bonds at the maleimide to the 4-thiovaleric acid terminus of the mertansine side chain and forms an amide bond between the carboxyl group of the linker and a lysine basic amine of the trastuzumab. Trastuzumab has 88 lysines (and 32 cysteines). As a result, trastuzumab emtansine is highly heterogeneous, containing dozens of different molecules containing from 0 to 8 mertansine units per trastuzumab, with an average mertansine/trastuzumab ratio of 3.4.

Suggestions



- Remove intolerance of Sgroup data in molfiles
- Support HELM-2 and SCSR as input formats
- Investigate performance issues
 - Canonicalization
 - Timeouts
- Enhance InChI data model to support
 - Variable substitution
 - Variable loading
 - Hydrogen bonds
 - Organometallic bonding
- Remove arbitrary limits
 - In particular maximum atom limit

Question



- Does InChI need to be a rigorous (valence-bond) representation of the structure?
- Is reproducible sufficient

Proposal



- Extend format with extra layers
 - Base InChI correlates to unsubstituted substance
 - Variable substructure
 - Loading variation – 1 to n
 - Position of loading
 - Use new flag to identify that InChI contains variable substituents and variable loading
- InChI key may need third section to contain variability information