

InChI for large molecules Workshop

Supported by: IUPAC Division VIII InChI subcommittee

NCBI/NLM

InChl 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 InChl 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 prerelease 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

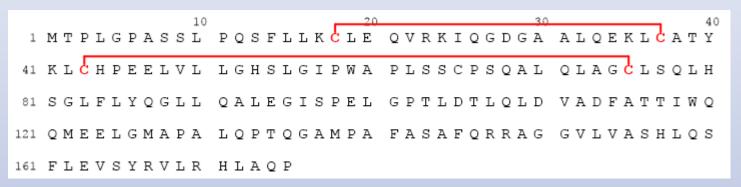
1	ŀ	1	т	Ρ	L	G	P	A	s	s	10 L	P	Q	S	F	L	L	ĸ	С	L	20 E	Q	v	R	K	I	Q	G	D	G	30 A	A	L	Q	E	K	L	С	A	т	40 Y
41	F	ζ	L	С	Н	Ρ	E	E	L	v	L	L	G	Н	s	L	G	I	P	W	A	Ρ	L	s	s	С	Ρ	s	Q	A	L	Q	L	A	G	С	L	s	Q	L	н
81	S	5	G	L	F	г	Y	Q	G	L	L	Q	A	L	E	G	I	s	Ρ	E	L	G	Ρ	т	L	D	т	L	Q	L	D	v	A	D	F	A	т	т	I	W	Q
121	ς	2	М	E	E	г	G	М	A	Ρ	A	L	Q	Ρ	т	Q	G	A	М	Ρ	A	F	A	s	A	F	Q	R	R	A	G	G	v	L	v	A	s	н	L	Q	s
161	E	7	L	E	V	s	Y	R	v	L	R	Н	L	A	Q	Ρ																									

InChIKey=KOKXRWZWQJXBOP-NJDFSSKJBA-N

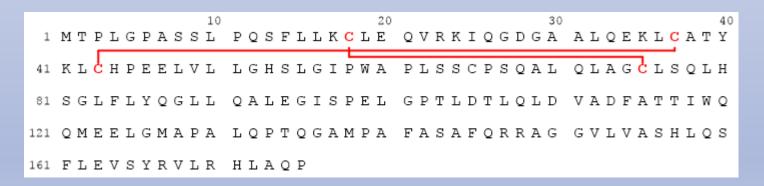
With disulfide bridges



InChIKey=MMCZGSMNPYTOPN-NJDFSSKJBA-N



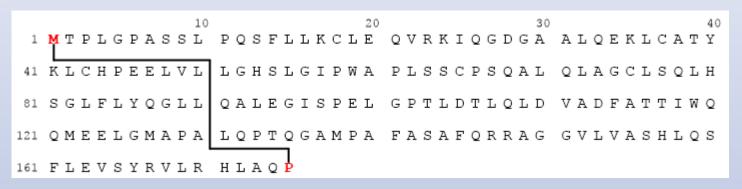
InChIKey=MEMBSBQMAVSGHQ-NJDFSSKJBA-N



Cyclized



• InChIKey=IZNXXFOUFDSLAX-VBNFVGOYBA-N



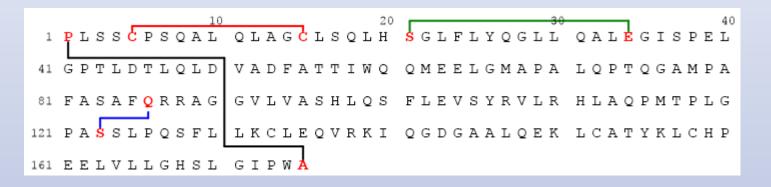
InChIKey=IZNXXFOUFDSLAX-VBNFVGOYBA-N

	10		20 30 40
1	PLSSCPSQAL	QLAGCLSQL	H SGLFLYQGLL QALEGISPEL
41	GPTLDTLQLD	VADFATTIW	Q QMEELGMAPA LQPTQGAMPA
81	FASAFQRRAG	GVLVASHLQ	S FLEVSYRVLR HLAQPMTPLG
121	PASSLPQSFL	LKCLEQVRK	I QGDGAALQEK LCATYKLCHP
161	EELVLLGHSL	GIPW A	

Multiple cyclizations



InChIKey=AQUGLJGKXYTOSD-VBNFVGOYBA-N



Reversed sequence



InChIKey=YFXNVYXMKDIHRN-VBNFVGOYBA-N

<mark>он</mark> 1	M	т	Ρ	L	G	Ρ	A	s	s	10 L	P	Q	s	F	L	г	ĸ	С	L	20 E	Q	v	R	ĸ	I	Q	G	D	G	30 A	A	L	Q	E	ĸ	L	С	A	т	40 Y
41	K	L	С	Н	Ρ	E	E	L	v	L	г	G	н	s	L	G	I	Ρ	W	A	Ρ	L	s	s	С	Ρ	s	Q	A	L	Q	L	A	G	С	L	s	Q	L	н
81	S	G	L	F	L	Y	Q	G	L	L	Q	A	L	E	G	I	s	Ρ	E	L	G	₽	т	L	D	т	L	Q	L	D	v	A	D	F	A	т	т	I	W	Q
121	Q	М	E	E	L	G	М	A	Ρ	A	L	Q	Ρ	т	Q	G	A	М	Ρ	A	F	A	s	A	F	Q	R	R	A	G	G	v	L	v	A	s	Н	L	Q	s
161	F	L	E	v	s	Y	R	v	L	R	Н	L	A	Q	Ρ	H																								

Filgrastim Lys10-D form



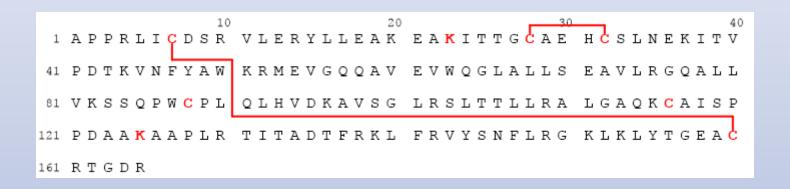
InChIKey=KOKXRWZWQJXBOP-FNWNWACTBA-N

	10	2	0 30	40
1	ΜΤΡLGPASS1	PQSFLLKCLE	QVRKIQGDGA ALQEKLCAT	Y
41	KLCHPEELVL	LGHSLGIPWA	PLSSCPSQAL QLAGCLSQL	н
81	SGLFLYQGLL	QALEGISPEI	GPTLDTLQLD VADFATTIW	Q
121	QMEELGMAPA	LQPTQGAMPA	FASAFQRRAG GVLVASHLQ	s
161	FLEVSYRVLR	HLAQP		

Synthetic Erythropoetin



InChIKey=XJBDLLBKVUYAKW-WAXLMBMOBA-D



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

Polynucleotide - 1



InChIKey=NELTZQNSFHRPGO-AZBJDUHQBA-N

	10	20	30	40	50
1	CGGAGCCTGC	AGCCCAGCCC	CACCCAGACC	CATGGCTGGA	CCTGCCACCC
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

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

Polynucleotide - 2

• InChIKey=IHDBOWIRPNDUCX-YUQJSOSJBA-N

	10	20	30	CATGGCTGGA	50
1	CGGAGCCTGC	AGCCCAGCCC	CACCCAGACC	CATGGCTGGA	CCTGCCACCC
51	AGAGCCCCAT	GAAGCTGATG	GCCCTGCAGC	TGCTGCTGTG	GCACAGTGCA
101	CTCTGGACAG	TGCAGGAAGC	CACCCCCTG	GGCCCTGCCA	GCTCCCTGCC
151	CCAGAGCTIC	CIGCICAAGI	GCTTAGAGCA	AGTGAGGAAG	ATCCAGGGCG
201	ATGGCGCAGC	GCTCCAGGAG	AAGCTGGTGA	GTGAGTGTGC	CACCTACAAG
251	CIGIGCCACC	CCGAGGAGCT	GGTGCTGCTC	GGACACTCTC	TGGGCATCCC
301	CIGGGCICCC	CTGAGCAGCT	GCCCCAGCCA	GGCCCTGCAG	CTGGCAGGCT
351	GCTTGAGCCA	ACTCCATAGC	GGCCTTTTCC	TCTACCAGGG	GCTCCTGCAG
401	GCCCTGGAAG	GGATCTCCCC	CGAGTTGGGT	CCCACCTTGG	ACACACTGCA
451	GCTGGACGTC	GCCGACTTTG	CCACCACCAT	CIGGCAGCAG	ATGGAAGAAC
501	TGGGAATGGC	CCCTGCCCTG	CAGCCCACCC	AGGGTGCCAT	GCCGGCCTTC
551	GCCTCTGCTT	TCCAGCGCCG	GGCAGGAGGG	GTCCTGGTTG	CCTCCCATCT

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



Polynucleotide - 3



InChlKey=

		10	20	30	40	50
	1	CGGAGCCTGC	AGCCCAGCCC	CACCCAGACC	CATGGCTGGA	CCTGCCACCC
	51	AGAGCCCCAT	GAAGCTGATG	GCCCTGCAGC	TGCTGCTGTG	GCACAGTGCA
	101	CTCTGGACAG	TGCAGGAAGC	CACCCCCTG	GGCCCTGCCA	GCTCCCTGCC
	151	CCAGAGCTTC	CTGCTCAAGT	GCTTAGAGCA	AGTGAGGAAG	ATCCAGGGCG
1	201	ATGGCGCAGC	GCTCCAGGAG	AAGCTGGTGA	GTGAGTGTGC	CACCTACAAG
1	251	CIGIGCCACC	CCGAGGAGCT	GGTGCTGCTC	GGACACTCTC	TGGGCATCCC
2	301	CIGGGCICCC	CTGAGCAGCT	GCCCCAGCCA	GGCCCTGCAG	CTGGCAGGCT
1	351	GCTTGAGCCA	ACTCCATAGC	GGCCTTTTCC	TCTACCAGGG	GCTCCTGCAG
1	901	GCCCTGGAAG	GGATCTCCCC	CGAGTTGGGT	CCCACCTTGG	ACACACTGCA
	\$51	GCTGGACGTC	GCCGACTTTG	CCACCACCAT	CIGGCAGCAG	ATGGAAGAAC
1	501	TGGGAATGGC	CCCTGCCCTG	CAGCCCACCC	AGGGTGCCAT	GCCGGCCTTC
1	551	GCCTCTGCTT	TCCAGCGCCG	GGCAGGAGGG	GTCCTGGTTG	CCTCCCATCT
	501	GCAGAGCTTC	CTGGAGGTGT	CGTACCGCGT	TCTACGCCAC	CTTGCCCAGC
	551	CCTGAGCCAA	GCCCTCCCCA	TCCCATGTAT	TTATCTCTAT	TTAATATTTA
1	701	TGTCTATTTA	AGCCTCATAT	TTAAAGACAG	GGAAGAGCAG	AACGGAGCCC
1	751	CAGGCCTCTG	TGTCCTTCCC	TGCATTTCTG	AGTTTCATTC	TCCTGCCTGT
	801	AGCAGTGAGA	AAAAGCTCCT	GTCCTCCCAT	CCCCTGGACT	GGGAGGTAGA
	951	TAGGTAAATA	CCAAGTATTT	ATTACTATGA	CTGCTCCCCA	GCCCTGGCTC

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



¹ MSSDSEMAIF GEAAPFLRKS²⁰ ERERIEAQNK PFDAKTSVF⁴⁰ VDPKESFVKA⁵⁰ ⁵¹ TVQSREGGKV TAKTEAGATV TVKDDQVFPM NPPKYDKIED MAMMTHLHEP 101 AVLYNLKERY AAWMIYTYSG LFCVTVNPYK WLPVYNAEVV TAYRGKKRQE ¹⁵¹ APPHIFSISD NAYOFMLTDR ENQSILITGE SGAGKTVNTK RVIQYFATIA 201 VTGEKKKEEV TSGKMQGTLE DQIISANPLL EAFGNAKTVR NDNSSRFGKF 251 IRIHFGTTGK LASADIETYL LEKSRVTFQL KAERSYHIFY QIMSNKKPDL 301 IEMLLITTNP YDYAFVSQGE ITVPSIDDQE ELMATDSAIE ILGFTSDERV 351 SIYKLTGAVM HYGNMKFKQK QREEQAEPDG TEVADKAAYL QNLNSADLLK JUL ALCYPRVKVG NEYVTKGQTV QQVYNAVGAL AKAVYDKMFL WMVTRINQQL ⁴⁵¹ DTKQPRQYFI GVLDIAGFEI FDFNSLEQLC INFTNEKLQQ FFNHHMFVLE ⁵⁰¹ QEEYKKEGIE WTFIDFGMDL AACIELIEKP MGIFSILEEE CMFPKATDTS ³⁰¹ FKNKLYEQHL GKSNNFQKPK PAKGKPEAHF SLIHYAGTVD YNIAGWLDKN ⁶⁰¹ KDPLNETVVG LYQKSAMKTL ALLFVGATGA EAEAGGGKKG GKKKGSSFQT ⁶⁵¹ VSALFRENLN KLMTNLRSTH PHFVRCIIPN ETKTPGAMEH ELVLHQLRCN 701 GVLEGIRICR KGFPSRILYA DFKQRYKVLN ASAIPEGQFI DSKKASEKLL 751 GSIDIDHTQY KFGHTKVFFK AGLLGLLEEM RDEKLAQLIT RTQAMCRGFL ⁸⁰¹ ARVEYQKMVE RRESIFCIQY NVRAFMNVKH WPWMKLYFKI KPLLKSAETE 801 KEMANMKEEF EKTKEELAKT EAKRKELEEK MVTLMQEKND LQLQVQAEAD 901 SLADAEERCD QLIKTKIQLE AKIKEVTERA EDEEEINAEL TAKKRKLEDE ⁹⁵¹ CSELKKDIDD LELTLAKVEK EKHATENKVK NLTEEMAGLD ETIAKLTKEK 1001 KALQEAHQQT LDDLQAEEDK VNTLTKAKIK LEQQVDDLEG SLEQEKKIRM 1051 DLERAKRKLE GDLKLAQEST MDIENDKQQL DEKLKKKEFE MSGLQSKIED 1101 EQALGMQLQK KIKELQARIE ELEEEIEAER ASRAKAEKQR SDLSRELEEI 1191 SERLEEAGGA TSAQIEMNKK REAEFQKMRR DLEEATLQHE ATAATLRKKH 1201 ADSVAELGEQ IDNLQRVKQK LEKEKSEMKM EIDDLASNME TVSKAKGNLE 1251 KMCRALEDQL SEIKTKEEEQ QRLINDLTAQ RARLQTESGE YSRQLDEKDT 1301 LVSQLSRGKQ AFTQQIEELK RQLEEEIKAK SALAHALQSS RHDCDLLREQ 1351 YEEEQEAKAE LQRAMSKANS EVAQWRTKYE TDAIQRTEEL EEAKKKLAQR 1401 LQDAEEHVEA VNAKCASLEK TKQRLQNEVE DLMIDVERTN AACAALDKKQ 1451 RNFDKILAEW KQKCEETHAE LEASQKESRS LSTELFKIKN AYEESLDQLE 1991 TLKRENKNLQ QEISDLTEQI AEGGKRIHEL EKIKKQVEQE KSELQAALEE 1551 AEASLEHEEG KILRIQLELN QVKSEVDRKI AEKDEEIDQM KRNHIRIVES 1601 MQSTLDAEIR SRNDAIRLKK KMEGDLNEME IQLNHANRMA AEALRNYRNT 1651 QAILKDTQLH LDDALRSQED LKEQLAMVER RANLLQAEIE ELRATLEQTE 1701 RSRKIAEQEL LDASERVQLL HTQNTSLINT KKKLETDISQ IQGEMEDIIQ 1751 EARNAEEKAK KAITDAAMMA EELKKEQDTS AHLERMKKNL EQTVKDLQHR 1801 LDEAEQLALK GGKKQIQKLE ARVRELEGEV ESEQKRNVEA VKGLRKHERK 1851 VKELTYQTEE DRKNILRLQD LVDKLQAKVK SYKRQAEEAE EQSNVNLSKF

InChIKey=BBJMARUZQDWUQG-PZLOAVSTBA-N

• Calculation time: ~94s

Myosin-1

• Molecular Formula: C₉₇₂₅H₁₅₈₁₆N₂₇₄₈O₃₁₀₀S₇₂

1901 RRIQHELEEA EERADIAESQ VNKLRVKSRE VHTKIISEE

Trastuzumab dimer



- InChIKey=VRBUFPXQWJVPLO-JNJMYDJTBA-N
- Single arbitrary stereocenter inverted
 - InChIKey=VRBUFPXQWJVPLO-RCEINSQCBA-N

- Calculation time: ~27s
- Molecular Formula: C₆₄₆₀H₉₉₇₂N₁₇₂₄O₂₀₁₄S₄₄

		10		20)	30		40		50	
	-		_	GGSLRL	SCAASG		DTYIHWV	_			
51	IYPTNGY	ſΤRΥ	ADSVE	KGRFTI	SADTSK	NTAY	LÖWNZTB	AED 1	AVYYCS	RWG	
101	GDGFYAN	1DYW	GOGTI	LVTVSS	ASTKGP	SVFP	LAPSSKS	TSG G	TAALGC	LVK	
151	DYFPEPV	JTVS	WNSGA	ALTSGV	HTFPAV	LQSS	GLYSLSS	VVT V	PSSSLG	TQT	
201	YICNVNE	IKPS	NTKVI	OKKVEP	KSCDKT	НТ <mark>С</mark> Р	PCPAPEL	LGG F	SVFLFP	DKD	
251	KDTLMIS	SRTP	EVIC	/VVDVS	HEDPEV	KENW	YVDGVEV	HNA B	TKPREE	QYN	
301	STYRVVS	SVLT	VLHQI	DWLNGK	EYKCKV	SNKA	LPAPIEK	TIS R	AKGQPR	EPQ	
351	VYTLPPS	SREE	MTKNC	DVSLTC	LVKGFY	PSDI	AVEWESN	GOP E	NNYKTT	PPV	
							MHEALHN	_			
						~~~ .					
1	DIOMTO	SPSS	LSAS	VGDRVŤ	ITCRAS	QDVN	TAVAWY	QKP (	GKAPKLI	LIYS	
51	ASFLYS	GVPS	RFSG	SRSGTD	FTLTIS	SLQE	EDFATY	CQQ I	HYTTPPI	FGQ	
							SVVCLL			-	
	DNALQS				-		LSKADY			-	
201	LSSPVTI	KSFN	RGEC	-	1					-	
				10	2	0	30		40		50
		_		_				DTYIN	_	PGKGLEWV	
	51	IYP?	CNCVTI	RY ADS	VKGRFTI	6 2 11					M C
			110111			SRD.	ISKNTAY	LOWNS	SLRAED	TAVYYCSR	MG
			FYAMD	YW GOG	TLVTVSS	AST	KGPSVFP	LAPSS	SKSTSG	<u>GTAALG</u> L	VK
			FYAMD	YW GOG	TLVTVSS	AST	KGPSVFP	LAPSS	SKSTSG		VK
	151	DYFI	FYAMD: PEPVTV	YW GOG VS WNS	TLVTVSS GALTSGV	ASTI HTFI	KGPSVFP PAVLQSS	LAPS: GLYSI	SKSTSG LSSVVT	<u>GTAALG</u> L	VK QT
	151 201	DYFI YI <b>CT</b>	FYAMD PEPVT NVNHKI	YW GOG VS WNS PS NTK	TLVTVSS GALTSGV VDKKVEP	ASTI HTFI KS <mark>C</mark> I	KGPSVFP PAVLQSS DKTHTCP	LAPS: GLYSI PCPAI	SKSTSG LSSVVT PELLGG	<u>GTAALGC</u> L VPSSSLGT	VK QT KP
	151 201 251	DYFI YI <mark>CI</mark> KDTI	FYAMD: PEPVTV NVNHKI LMISR:	YW GOG VS WNS PS NTK TP EVT	TLVTVSS GALTSGV VDKKVEP CVVVDVS	ASTI HTFI KSCI HEDI	KGPSVFP PAVLQSS DKTHTCP PEVKFNW	LAPS GLYSI PCPAI YVDGV	SKSTSG LSSVVT PELLGG VEVHNA	GTAALG <mark>C</mark> L VPSSSLGT PSVFLFPP	VK QT KP YN
	151 201 251 301	DYFI YICT KDTI STYF	FYAMD PEPVTV NVNHKI LMISR RVVSVI	YW GOG VS WNS PS NTK IP EVT LT VLH	TLVTVSS GALTSGV VDKKVEP CVVVDVS QDWLNGK	ASTI HTFI KSCI HEDI EYK	KGPSVFP PAVLQSS DKTHTCP PEVKFNW CKVSNKA	LAPSS GLYSI PCPAI YVDGV LPAPS	SKSTSG LSSVVT PELLGG /EVHNA IEKTIS	<u>GTAALGC</u> L VPSSSLGT PSVFLFPP KTKPREEQ	VK QT KP YN PQ
	151 201 251 301 351	DYFI YICT KDTI STYF VYTI	FYAMD PEPVTV VVNHKT LMISR RVVSVI LPPSRI	YW GOG VS WNS PS NTK TP EVT LT VLH EE MTK	TLVTVSS GALTSGV VDKKVEP CVVVDVS QDWLNGR NQVSLTC	ASTI HTFI KSCI HEDI EYK(	KGPSVFP PAVLQSS DKTHTCP PEVKFNW CKVSNKA GFYPSDI	LAPS: GLYSI PCPAI YVDGV LPAP: AVEWI	SKSTSG LSSVVT PELLGG VEVHNA IEKTIS SNGQP	GTAALGCL VPSSSLGT PSVFLFPP KTKPREEQ KAKGQPRE	VK QT KP YN PQ PV
	151 201 251 301 351	DYFI YICT KDTI STYF VYTI	FYAMD PEPVTV VVNHKT LMISR RVVSVI LPPSRI	YW GOG VS WNS PS NTK TP EVT LT VLH EE MTK	TLVTVSS GALTSGV VDKKVEP CVVVDVS QDWLNGR NQVSLTC	ASTI HTFI KSCI HEDI EYK(	KGPSVFP PAVLQSS DKTHTCP PEVKFNW CKVSNKA GFYPSDI	LAPS: GLYSI PCPAI YVDGV LPAP: AVEWI	SKSTSG LSSVVT PELLGG VEVHNA IEKTIS SNGQP	GTAALGC VPSSSLGT PSVFLFPP KTKPREEQ KAKGQPRE ENNYKTTP	VK QT KP YN PQ GK
	151 201 251 301 351 401	DYFI YICT KDTI STYF VYTI LDSI	FYAMD PEPVTV NVNHKI LMISR RVVSVI LPPSRI DGSFFI	YW GOG VS WNS PS NTK TP EVT LT VLH EE MTK LY SKL	TLVTVSS GALTSGV VDKKVEP CVVVDVS QDWLNGK NQVSLTC TVDKSRW	ASTI HTFI KSCI HEDI EYKU LVK( QQGI	KGPSVFP PAVLQSS DKTHTCP PEVKFNW CKVSNKA GFYPSDI NVFSCSV 30	LAPS GLYSI PCPAI YVDGV LPAP AVEWI MHEAI	SKSTSG LSSVVT PELLGG VEVHNA IEKTIS ESNGQP LHNHYT	GTAALGC VPSSSLGT PSVFLFPP KTKPREEQ KAKGQPRE ENNYKTTP QKSLSLSP	VK QT KP YN PQ GK 50
	151 201 251 301 351 401	DYFI YICT KDTI STYP VYTI LDSI DIQI	FYAMD PEPVTV NVNHKI LMISR RVVSVI LPPSRI DGSFFI MTQSP	YW GOG VS WNS PS NTK IP EVT LT VLH EE MTK LY SKL S ¹⁰ LSA	TLVTVSS GALTSGV VDKKVEP CVVVDVS QDWLNGK NQVSLTC TVDKSRØ SVGDRV	ASTI HTFI KSCI HEDI EYKO LVKO QQGI ITC	KGPSVFP PAVLQSS DKTHTCP PEVKFNW CKVSNKA GFYPSDI NVFSCSV RASQDVN	LAPS GLYSI PCPAI YVDGV LPAP: AVEWI MHEAI TAVA	SKSTSG LSSVVT PELLGG VEVHNA IEKTIS ESNGQP LHNHYT WYQQKP	GTAALGC VPSSSLGT PSVFLFPP KTKPREEQ KAKGQPRE ENNYKTTP QKSLSLSP GKAPKLLI	VK QT XP YN PQ GK XS
	151 201 251 301 351 401 1 51	DYFI YICT KDTI STYF VYTI LDSI DIQI ASF	FYAMD PEPVTV NVNHKI LMISR RVVSVI LPPSRI DGSFFI MTQSP LYSGV	YW GOG VS WNS PS NTK IP EVT LT VLH EE MTK LY SKL S ¹⁰ LSA PS RFS	TLVTVSS GALTSGV VDKKVEP CVVVDVS QDWLNGK NQVSLTC TVDKSRØ SVGDRV GSRSGTI	ASTI HTFI KSCI HEDI EYKO LVKO QQGI ITC FTL	KGPSVFP PAVLQSS DKTHTCP PEVKFNW CKVSNKA GFYPSDI NVFSCSV RASQDVN TISSLQP	LAPSS GLYSI FCPAI YVDGV LPAP: AVEWI MHEAI TAVA	SKSTSG LSSVVT PELLGG JEVHNA IEKTIS ESNGQP LHNHYT WYQQKP TYYCQQ	GTAALGCL VPSSSLGT PSVFLFPP KTKPREEQ KAKGQPRE ENNYKTTP QKSLSLSP GKAPKLLI HYTTPPTF	VK QT XP PQ GK GK
	151 201 251 301 351 401 1 51	DYFI YICT KDTI STYF VYTI LDSI DIQI ASF	FYAMD: PEPVTV NVNHKI LMISR: RVVSVI LPPSRI DGSFFI MTQSP LYSGV VEIKR	YW GOG VS WNS PS NTK IP EVT LT VLH EE MTK LY SKL S ¹⁰ LSA PS RFS TV AAP	TLVTVSS GALTSGV VDKKVEP CVVVDVS QDWLNGK NQVSLTC TVDKSRØ SVGDRV GSRSGTI SVFIFPI	ASTI HTFI KSCI HEDI EYKC LVK( QQGI ITC FTL SDE	KGPSVFP PAVLQSS DKTHTCP PEVKFNW CKVSNKA GFYPSDI NVFSCSV RASQDVN TISSLQP QLKSGTA	LAPSS GLYSI PCPAI YVDGV LPAP: AVEWH MHEAD TAVA EDFA SVVC	SKSTSG LSSVVT PELLGG VEVHNA LEKTIS ESNGQP LHNHYT WYQQKP TYYCQQ LLNNFY	GTAALGCL VPSSSLGT PSVFLFPP KTKPREEQ KAKGQPRE ENNYKTTP QKSLSLSP GKAPKLLI HYTTPPTF PREAKVQW	VK QT XP PQ GK GK GQ KV
	151 201 301 351 401 1 51 101	DYFI YICT KDTI STYF VYTI LDSI DIQI ASF GTK DNA	FYAMD' PEPVTV NVNHKI LMISR' RVVSVI LPPSRI DGSFFI MTQSP LYSGV VEIKR LQSGN	YW GOG VS WNS PS NTK IP EVT LT VLH EE MTK LY SKL S ¹⁰ LSA PS RFS TV AAP	TLVTVSS GALTSGV VDKKVEP CVVVDVS QDWLNGK NQVSLTC TVDKSRØ SVGDRV GSRSGTI SVFIFPI TEQDSKI	ASTI HTFI KSCI HEDI EYKC LVK( QQGI ITC FTL SDE	KGPSVFP PAVLQSS DKTHTCP PEVKFNW CKVSNKA GFYPSDI NVFSCSV RASQDVN TISSLQP QLKSGTA	LAPSS GLYSI PCPAI YVDGV LPAP: AVEWH MHEAD TAVA EDFA SVVC	SKSTSG LSSVVT PELLGG VEVHNA LEKTIS ESNGQP LHNHYT WYQQKP TYYCQQ LLNNFY	GTAALGCL VPSSSLGT PSVFLFPP KTKPREEQ KAKGQPRE ENNYKTTP QKSLSLSP GKAPKLLI HYTTPPTF	VK QT XP PQ GK GK GQ KV

# Summary



- InChl v1.05 can generate InChl 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