Comments on polymer InChI issues

There are two issues on InChI for polymers which were recently debated;   
they were mentioned by John Mayfield (October 2016, discussion: https://sourceforge.net/p/inchi/mailman/message/35421506/) and then were re-iterated in the NextMove presentation by R.Sayle, J.Mayfield, N.O’Boyle (<http://www.slideshare.net/NextMoveSoftware/rdkit-six-notsoeasy-pieces-rdkit-ugm-2016>) and then followed by Evan Bolton on late January 2017 (e-mail exchange).

*#1 No reduction of SRU (structural repeating unit of polymer) to simplest possible one is performed*

*#2 No phase shift (frame shift) of SRU is performed unless endgroups are abstract star atoms '\*'*

In short: these features are known and intentional (note that issue #2 is explicitly described in documentation). The related issues were not considered critical - at least for the first (beta) stage of implementing InChI for polymers.

Details and comments are as follow.

1. What is the problem

*Issue #1 No reduction of SRU to simplest possible one is performed*

It is about occurrence of repeating unit within repeating unit:   
consider polymethylene vs. polyethylene or, in general, -[-X-]n- *vs.* -[-X-X-]n-, -[-X-X-X-]n- and so on. These structures currently receive different InChI’S/Key’s:



*Issue #2 No phase shift (frame shift) of SRU (structural repeating unit of polymer) is performed unless endgroups are abstract star atoms '\*'*

Current implementation does recognize "phase shift" \_only\_ if polymeric end groups are represented by star atoms.

That is, \*-based polymers

\* -[-C-C(O)-NH-]n-\* \*-[-C(O)-NH-C-]n-\* \*-[-NH-C-C(O)-]n-\*

all receive the same InChI/Key's (as desirable):



while indicating explicit end groups kills the magics – identifiers become different:



1. Why it is so

*Issue #1*

The issue of “SRU simplification“ was not mentioned by experts of Polymers task group in their Guidelines document. So it was simply considered out of scope -- or overlooked, if one prefers saying so.

Additional note is that, in current implementation, care was taken to to avoid, by all means, highly "invasive" operations like re-drawing input structure which changes the empirical formula; note that SRU simplification is such invasive. Note also that all "normalization" operations previously allowed in InChI were not more severe than adding/removing protons or minor changing the bonding patterns.

*Issue #2*

Guidelines document of Polymers task group did suggest that phase shift in non-star-capped SRU polymers is considered and accounted for.

Quote: “*End-groups are dependent on representation of CRU and for cononicalization of CRU it is necessary to modify end-groups to be in accord with possible connections in a polymer. ... Note that canonicalization for such presentations is more complex as soon as needs to take into account end-groups and retain correct connection of CRUs.*”

However, the limitation by issue #2, for this release, was introduced -- as a result of balancing of its importance vs. effort needed to implement, relating to this specific case of SRU canonicalization. A partial justification is that, as stated in the Guidelines: “*In most cases end-groups in polymer are not principle and are not specified.*... *Specification of end-groups in copolymers is rarely used*...”

Note that this limitation is explicitly documented.

More details.

Please note that in those 3 examples above (the same as on slide 21 of NextMove presentation) the capping groups in pGly\_1, pGly\_2, pGly\_3 are all different. The first part of InChI/Key is the same for all 3 drawings, and the last (polymer) layer is different, reflecting different placing of polymer brackets. Of course, ideally all the 3 InChI/Key's would be the same... and SRU - which is drawn in 3 cases as seemingly different, due to frame shift - would be encoded the same... and this is indeed ensured -- but only if the end groups are all stars, '\*'

However, making this behavior general, independent on changing capping groups, is not easy, I am afraid. I more or less understand how to make this behavior work for simple examples with explicit end groups like those on Slide 21... but the things quickly become much more complex for general case, where, e.g., end groups may contain their own nested repeating units. In view of this, it was decided to limit phase-shift (frame shift) analysis by '\*'-capped polymers and explicitly state this, as a lesser evil.

1. What to do

Below are the comments on possible resolving/improvement with the issues.

*#1 SRU simplifcation*

This simplification is doable - with the two reservations:

1. Simplification ( -XX- to -X- or alike ) may appear not so easy for complex X/environment.

2. Even after simplification InChI strings for different repeats will not be exactly the same.

Note that drawing -[-XX-]n- is nearly but not totally the same as -[-X-]n-. Namely, the former may be simplified to the latter with supposition that the latter's number of repeats 'n' is even. Simply drawing -[-X-]n-covers both odd and even number of repeats. Analogously, this is applied for -[XXX]-n etc. So the simplified SRU's will be the same, but number of 'inner-repeats' must be encoded some way; of course, the only right place is polymer layer. That is, -XX- may indeed be simplified to -X-, but resulting InChI will differ (from that of plain -X-) in polymer layer. Still, as the first parts of InChI and InChIKey, up to /z layer, will be the same, it gives more convenience than we have currently.

*#2 Canonicalization of SRU in case of explicit (non-star) end groups*

It is also doable, but here I have more doubts. I more or less understand how to achieve this canonicalization in simple cases like polyglicine above.. and definitely expect difficulties for more complex cases.

1. Conclusion

My opinion is that this discussion should not affect current v. 1.05 release.

For the (near) future, I guess we should carefully balance an actual importance of resolving above mentioned issues (features) vs. the necessary effort, given our limited possibilities.

The result of this balancing is not evident for me.

I suggest that we ask more feedback than we currently have -- I guess not through messaging in inchi-discuss or alike forums but through direct addressing the specialists interested/proficient in polymer chemistry.