Mission and challenges of Polymer Informatics

Preeti Chhokkar

Department of Physics, Kurukshetra University Kurukshetra, Haryana, India

Abstract- Polymers are arguably the most important set of materials in common use. The increasing adoption of both combinatorial as well as high-throughput approaches, coupled with an increasing amount of interdisciplinarity, has wrought tremendous change in the field of polymer science. Yet the informatics tools required to support and further enhance these changes are almost completely absent. In the first part of the chapter, a critical analysis of the challenges facing modern polymer informatics is provided. It is argued, that most of the problems facing the field today are rooted in the current scholarly communication process and the way in which chemists and polymer scientists handle and publish data. Furthermore, the chapter reviews existing modes of representing and communicating polymer information and discusses the impact, which the emergence of semantic technologies will have on the way in which scientific and polymer data is published and transmitted. In the second part, a review of the use of informatics tools for the prediction of polymer properties and in silico design of polymers is offered..

Keywords- polymer informatics, polymer markup language, chemical markup language, polymer ontology

I. INTRODUCTION

Polymer informatics combines polymer chemistry, computer science and information science. The idea of polymer informatics is to advance the design, analysis and understanding of polymer systems. A polymer informatician probes and employs insights from the systematic study of computational methods, knowledge acquisition strategies and pattern recognition algorithms to develop digitalized solutions for polymer research & engineering.

Like the related disciplines cheminformatics and bioinformatics, polymer informatics is an interdisciplinary field. It is an emerging discipline that should not be considered a subdiscipline of cheminformatics. Cheminformatics "deals" with small molecules, i.e. molecules with a confined structure whose composition and atom connectivity can precisely be represented by a molecular graph and an associated connection table. The subject of polymer informatics is the rational management of macromolecules—chain-like molecules consisting of one or more structural repeat units (SRUs). Regular single- and multi-strand polymers and copolymers are the key ingredients of polymer systems; for example blends and composites. Cheminformatics and polymer informatics are mostly design-oriented. In contrast, bioinformatics pays particular attention to the sequence patterns (typically nucleic acid and protein sequences) of biomacromolecules within the context of biological processes and gene-based drug discovery.

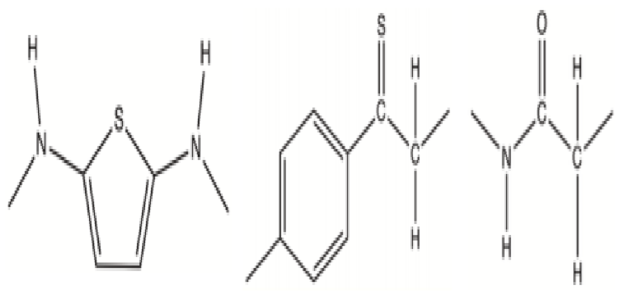

Figure 1 Typical fragments

Critical for the unambiguous description, storage, search and modeling of polymer systems is the adoption of recommended, agreed-upon nomenclatures and structural representation systems. An IUPAC recommendation for organic polymers exists and provides a structure-based nomenclature for regular single-strand polymers [1]. The chemical Sgroup approach serves as a polymer abstraction concept [2].

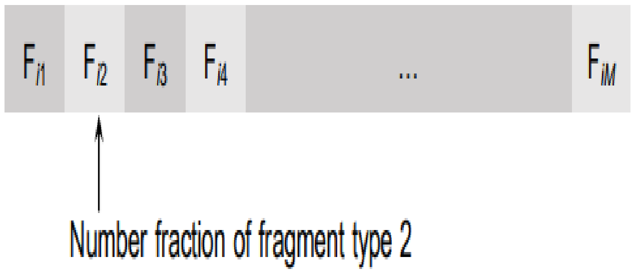

Figure 2 Schematic of organic polymer fingerprint construction

The Polymer Markup Language (PLM) utilizes XML technology to manage polymer information [3]. The user-friendly CurlySMILES language supports structural encoding of macromolecules as annotated SMILES notation [4,5], CurlySMILES is currently enhanced for the encoding of multi-stand polymers and copolymers. Further, CurlySMILES provides a syntax to represent complex systems such as polymer assemblies, polymer solutions, doped polymers and nanocomposites in a compact single line notation. A recent thesis on automatic polymer data evaluation in combination with the Polymer Informatics Knowledge System (PIKS) constitutes an excellent source to familiarize oneself with solutions and challenges in computer-assisted polymer research [6]. The present Polymer Informatics blog is intended as a platform to discuss diverse aspects of integrating polymer science with data management technologies and computational disciplines

Figure 1: Learning polymer properties using fragment-level fingerprints.

II. NATURE OF POLYMER INFORMATION

Small molecule informatics is in essence a solved problem. A number of methods and technologies exist to represent molecules to a machine in multiple dimensions (0 - 3D), ranging from trivial and systematic names and brutto formulae to line notations such as the “simplified molecular input line entry specification”[23] (SMILES) and the International Union of Pure and Applied Chemistry’s (IUPAC) International Chemical Identifier[24] (InChI) and to full connection tables in a plethora of formats, such as mol, pdb or Chemical Markup Language [25-28] (CML). These representations are normally constructed on the basis of results derived from modern analytical chemistry, which can be successfully used to elucidate the structure and therefore the “connection table” of small molecules. While chemists are accustomed to think of both small molecules and polymers as “substances”, i.e. a particular kind of matter with uniform properties, there is a profound difference between the two, which causes confusion and difficulties for the chemical information scientist. Unlike substances composed of well-defined small molecules of usually identical structure, polymers consist of ensembles of macromolecules, all of which have slightly different architectures (in the simplest case only differing by length, in more complicated cases showing extensive branching or cross-linking) and therefore slightly different properties.[29] Physical quantities commonly referred to as “polymer properties” do not relate to a pure substance with a unique connection table, but are averages over structurally diverse ensembles of macromolecules. Molecular weight distributions in classically prepared synthetic 8 polymers are unavoidable – even the most controlled polymerisations lead to polydispersity indices (PDIs) larger than 1 (very controlled living polymerisations achieve PDIs of around 1.03 (see, for example, reference [30]). Furthermore, even modern analytical tools do not allow for the “connection table” of all of the constituent macromolecules in an ensemble to be determined, which makes the accurate description of a polymer in terms of the structures of its constituent macromolecules impossible and introduces a significant fuzziness of concept. The latter, in turn, breaks the transition from structure to property, which traditional chemical informatics is trying to make.

III. REPRESENTATION OF POLYMERS

The fuzziness of concept discussed above can be found right across polymer science and probably nowhere more so than in the representation of polymers to machines (e.g. in databases etc.). Typically, polymers are represented in information systems using either a name (a text string) or an idealised/abstracted or reduced structural description (an idealised connection table, a graphical representation) or a combination of both. Both types of representations have their particular problems.

Name-based representations.

Name-based representations are normally constructed either from the component monomers of a polymer (source-based representations) or from the repeating unit (structure-based representation) and frequently trivial names are still in use. Each of these representations has merits and disadvantages and there is no general agreement in the polymer science community, as to which representation is preferable. Furthermore, the form which the name based representation will take, depends on the 9 different nomenclature philosophies used across chemistry. As an example, consider the representation of the polymer with the repeat unit structure depicted in Figure 2. The Chemical Abstracts Service (CAS) will register the polymer as “1,3-butadiene, homopolymer”[31] whereas IUPAC allows the use of “polybutadiene” (IUPAC source based), “poly(but-1-ene-1,4-diyl)” (IUPAC structure based), “1,4-polybutadiene” (IUPAC semisystematic name) or “poly(buta-1,3-diene)” (IUPAC source based).[32] In addition to the different representation conventions (source-based/structure-based), these examples also illustrate the inversion of names for registration purposes (CAS), as well as the inconsistent use of brackets. Furthermore, each nomenclature and registration system has its own historical continuity - as the system evolves, naming conventions and therefore registrations change. The CAS 8th collective index (CI) name for poly(ethylene terephthalate) (Figure 3), for example, is poly(oxyethyleneoxyterephthaloyl), whereas the 9th CI name is poly(oxy-1,2-ethanediyloxycarbonyl-1,4-phenylenecarbonyl) (at the time of writing, Chemical Abstracts is in the 15th CI period). However, many chemists continue to use old nomenclature or even trivial names in their daily work: “methyl methacrylate” is still the preferred representation for a particular monomer molecule, rather than “methacrylic acid, methyl ester” (8th CI) or even “2-propenoic acid, 2-methyl-, methyl ester” (9th CI). It is not merely enough for rules and conventions to exist and to be implemented in a closed system such as the Chemical Abstracts: they also need to be adopted by a significant number of practicing chemists to be useful. While the plethora and complexity of possible name-based representations may, at worst, be confusing to the human chemist, it causes significant problems for the information scientist and the computer. Firstly, it may lead to multiple registrations of the same compound in a database, which, in turn, often results in only partial retrieval 10 of information associated with the same concept: unless one remembers to search for polybutadiene as well as all other possible representations of the same substance (taking into account both synonyms and historical continuity), one may not all the desired information. Even more gravely, the scenario outlined above requires a software agent to retrieve information about a polymer from different sources (e.g. physico-chemical properties database, toxicology database) and to subsequently unify the information. The unification process is essentially a mapping procedure, which requires software to recognize concepts as equivalent: while a chemist may be able to recognize, that the labels “poly(but-1-ene-1,4-diyl)” and “poly(buta-1,3-diene)” refer to equivalent concepts, this would be impossible for a machine if it had to exclusively rely on name based representations alone.

Graphical representations.

An idealized or abstracted structural sketch can also be used to represent polymers. “Structural” in this context refers to the use of chemical structure diagrams as a graphical metaphor for a connection table and should not be confused with the structure-based representations discussed above. When examining the polymer shown in Figure 4, it becomes evident that several valid repeat unit structures can be drawn (the possible repeat units A, B and C are “phase-shifted” with respect to each other) and therefore no unambiguous definition of a representation is possible in the absence of further specifying guidelines. In order to determine the preferred representation, a set of rules has to be developed and adopted by the chemical community. IUPAC defines an elaborate set of rules based on seniority of subunits, the “direction of citation” etc..[33] In this context, it is important to remember, that although we are discussing the choice of the preferred repeat unit in terms of a graphical 11 representation, these rules also influence the construction of polymer names, where the name is structure-based. Further rules are used to refine these constructs. From the point of view of an information scientist, this raises problems similar to the ones discussed for name-based representations: the rules governing a rule-based system must be accepted and followed if a consistent and unambiguous representation of polymers is to be achieved. Each of these systems, however, also exists in time and is therefore subject to change, which introduces added layers of complexity. The complexity is further increased, when several competing nomenclature systems are available, which essentially multiply the problems discussed so far. The discussion presented here has only focussed on simple linear polymers and even for those it has barely scratched the surface. Nomenclature and registration systems for polymers have been extensively reviewed by Wilks and others and the reader is referred to the literature for further information.[32,34-38] A paper, published in the early 1990s commented that “Just the mention of the word “polymer” has been known to strike fear into the hearts of mere mortals and certainly, at the least, a sense of apprehension, if not foreboding to an information researcher.” [37] Sadly, the situation has not changed significantly over the last decade.

IV. SOURCES OF POLYMER INFORMATION

In a set of introductory remarks at an ACS symposium on the retrieval of polymer information, Metanomski remarked in the late 1970ies, that it “is extremely important to have an easy and reliable access to the numerical data (preferably evaluated and verified) as well as to a variety of properties […].”[39] The two main concepts in this remark, namely “access” and “evaluated/verified data” remain as pressing and unfortunately unaddressed as they were almost two decades ago. 12 2.1.2.1 Access We have already discussed the fact, that polymer science is becoming increasingly data-centric, with high-throughput and combinatorial approaches being adopted as main-stream tools in the laboratory. However, the way in which science has chosen to report and archive its results generally leads to fragmentation, inaccessibility and the development of knowledge silos. The majority of polymer (-related) data originates from a small number of sources, namely scientific publications, theses and data compilations. In order to be able to extract data and mine these sources, they first need to be accessed by a machine. There are a number of obstacles to access, such as the physical availability of data (is it available electronically or as a paper copy on a library shelf, non-destructive document formats and copyright considerations. The requirement for the electronic availability of data and documents is obvious, if a software agent is to discover information. Although more and more institutions now require theses and dissertations to be reposited as a condition of granting a degree, this is still far from universal and a significant number are archived on a paper-only basis by libraries. However, even if available electronically, the format, in which the document is available, is critical. Most science papers and theses are either authored in LaTeX[40] or other text processing systems such as Microsoft Word[41] or Open Office and are subsequently – more often than not - converted to portable document format (pdf) for printing, distribution and repositing. The conversion to pdf, however, often destroys vital scientific information: the process converts text to a set of graphical objects without semantics, i.e. without well-defined relationships between them. For example, the information concerning superscripts and subscripts (which could identify chemical 13 formulae) is lost. Furthermore, the resulting graphical objects, cannot be processed further by computers in a data extraction/mining exercise and have to be converted back to text. As, at this stage, a significant amount of important information has been destroyed during the initial conversion process, the back-conversion yields unsatisfactory results such as jumbled data tables and formulae, which are difficult to interpret for both human and machine (Figure 5). In the context of our vision for polymer informatics, in which a software programme automatically detects and gathers data and information, this clearly presents a major obstacle. The most machine-friendly ways of transmitting and storing information is plain text, which is augmented with a form of text-based markup (such as LaTeX, HTML and XML documents), as information transmission here is usually lossless. Furthermore, closed proprietary formats also present problems for long-term storage and archival, particularly if the software required to access them, no longer exists.[42] Beyond these more technical considerations, the structure of a document also needs to be taken into account when considering access to data. The main form of communication in the chemical sciences is the scientific paper (and to a lesser extent the thesis), which typically intersperses (polymer) data with free text, thus effectively forming a "datument," (data + document) albeit an unstructured one.[43] It is difficult for a machine to automatically discover chemical information in collections of unstructured documents, as these are inevitably semantically poor. A typical example of a sentence that could be found in an unstructured datument could be: "poly(styrene) has a glass transition temperature of 99 °C". Without the availability of structuring metadata or a significant amount of "information archaeology", a machine has little chance to discover that the concept "poly(styrene)" refers to a polymer and "glass transition temperature" to a polymer property which, in turn, usually has an 14 associated value and a unit. If, however, concepts, values and units could be marked up as such in a machine discoverable way, this information could be extracted and made available for further processing. Markup of this type as part of the text would convert the unstructured datument to a fully structured one

V. CONCLUSIONS

Polymers are arguably the most important set of materials in common use. The increasing adoption of both combinatorial as well as high-throughput approaches, coupled with an increasing amount of interdisciplinarity, has wrought tremendous change in the field of polymer science. Yet the informatics tools required to support and further enhance these changes are almost completely absent. In the first part of the chapter, a critical analysis of the challenges facing modern polymer informatics is provided. It is argued, that most of the problems facing the field today are rooted in the current scholarly communication process and the way in which chemists and polymer scientists handle and publish data. Furthermore, the chapter reviews existing modes of representing and communicating polymer information and discusses the impact, which the emergence of semantic technologies will have on the way in which scientific and polymer data is published and transmitted. In the second part, a review of the use of informatics tools for the prediction of polymer properties and in silico design of polymers is offered. KeywordsInformation systems-Machine learning-Ontology-Polymer-markup language-Polymer informatics-QSPR-RDF-Semantic web.

VI. REFERENCES

[1]. Yoshida M, Langer R, Lendlein A et al. (2006) From advanced biomedical coatings to multi-functionalized biomaterials. Polym Rev 46:347–375

[2]. Dewez JL, Lhoest JB, Detrait E et al. (1998) Adhesion of mammalian cells to polymer surfaces: from physical chemistry of surfaces to selective adhesion on defined patterns. Biomaterials 19:1441–1445

[3]. Brocchini S, James K, Tangpasuthadol V et al. (1998) Structure-property coorrelations in a combinatorial library of degradable biomaterials. J Biomed Mat Res 42:66–75

[4]. Cuchelkar V, Kopecek J (2006) Polymer-drug conjugates. In: Uchegbu IF and Schaetzlein AG (ed) Polym Drug Deliv, CRC Press, Boca Raton

[5]. Torchilin VP (2006) Polymorphic micelles as pharmaceutical carriers. Polym Drug Deliv 111–130

[6]. Haag R, Kratz F (2006) Polymer therapeutics: concepts and applications. Angew Chem Int Edn 45:1198–1215

[7]. Khandare J, Minko T (2006) Polymer-drug conjugates: progress in polymeric prodrugs. Prog Polym Sci 31:359–397

[8]. Way JL, Petrikovics I, Jiang J et al. (2001) Application of dendrimeric polymers as a drug carrier in pharmacology. Abstracts of Papers, 221st ACS National Meeting, San Diego, CA, United States, April 1–5, 2001 IEC-316

[9]. Kataoka K, Kwon GS, Yokoyama M et al. (1993) Block copolymer micelles as vehicles for drug delivery. J Contr Rel 24:119–132

[10]. Malmsten M (2006) Soft drug delivery systems. Soft Matter 2:760–769

[11]. Qiu LY, Bae YH (2006) Polymer architecture and drug delivery. Pharm Res 23:1–30

[12]. Kang HC, Lee M, Bae YH (2007) Polymeric gene delivery vectors. In: Peppas NA, Hilt JZ, Thomas JB (ed) Nanotechnology in therapeutics Taylor and Francis, New York

[13]. Alexis F, Zeng J, Wang S (2007) PEI nanoparticles for targeted gene delivery. Gene Transfer 473–478

[14]. Leong KW (2006) Polymer design for nonviral gene delivery. BioMEMS Biomed Nanotechnol 1:239–263

[15]. Mahato RI (2005) Water insoluble and soluble lipids for gene delivery. Adv Drug Deliv Rev 57:699–712

[16]. Mahato RI, Kim SW (2005) Water soluble lipopolymers for gene delivery. In: Ammiji MM (ed) Polym Gene Deliv, CRC Press, Boca Raton

[17]. Adams ML, Lavasanifar A, Kwon GS (2003) Amphiphilic block copolymers for drug delivery. J Pharm Sci 92:1343–1355

[18]. Wagner E, Kloeckner J (2006) Gene delivery using polymer therapeutics. Adv Polym Sci 192:135–173

[19]. Joester D, Losson M, Pugin R et al. (2003) Amphiphilic dendrimers: novel self-assembling vectors for efficient gene delivery. Angew Chem Int Ed 42:1486–1490

[20]. Bjornerg HC, Derici L, Haggman BH et al. (2006) Hair care compositions comprising a dendritic polymer. 2005-EP7017 2006018064

[21]. Derici L, Harcup JP, Khoshdel E (2006) Hair care composition comprising a dendritic macromolecule. 2005-EP7016 2006018063

[22]. Goosey M (2007) An overview of polymers as key enablers in electronics assembly-a printed circuit board perspective. Polymers in Electronics 2007: Paper9/1-Paper9/5, Munich, Germany

[23]. Rost H (2007) Printed electronic circuits. Kunstst 97:97–101

[24]. Xing R-b, Ding Y, Han Y-c (2007) Patterning of polymer by inkjet printing and its application in the fabrication of organic electronic devices. Fenzi Kexue Xuebao 23:75–81

[25]. Liang Z, Wang Q (2007) Patterning of conjugated polymers for organic electronics and optoelectronics. In: Naiwa HS (ed) Polym Nanostruct Their Appl, American Scientific Publishers, Stevenson Ranch, California.