**Proceedings of the Conference on
Promoting Undergraduate Research in Mathematics**

REU Site: Algorithmic Combinatorics on Words

Francine Blanchet-Sadri

a. Overview

The intellectual focus of this "Research Experiences for Undergraduates (REU)" NSF supported program entitled *Algorithmic Combinatorics on Words* is on interdisciplinary research at the crossroads between Mathematics and Computer Science. *Combinatorics on words* is a rather new field although the first papers were published at the beginning of the 20th century. It grew independently in various areas of mathematics including group theory and number theory, and appears frequently in problems related to automata and formal language theory. In the latest classification of Mathematical Reviews, combinatorics on words constitutes its own section under discrete mathematics related to computer science.

Molecular biology has stimulated considerable interest in the study of *partial words* which are strings that may contain a number of "do not know" symbols or "holes". The motivation behind the notion of a partial word is the comparison of genes. Alignment of two such strings can be viewed as a construction of two partial words that are said to be compatible. While a word can be described by a total function, a partial word can be described by a partial function. More precisely, a partial word of length n over a finite alphabet Σ is a partial function from $\{0,\ldots,n-1\}$ into Σ. Elements of $\{0,\ldots,n-1\}$ without an image are called holes (a word is just a partial word without holes). Research in algorithmic combinatorics on partial words is underway and has the potential for impacts in numerous areas, notably in molecular biology, nano-technology, and DNA computing.

Since the summer of 2005, the University of North Carolina at Greensboro (UNCG) has provided unique opportunities for summer research for eight outstanding and highly motivated students per year for an eigth-week period each year (note that in Summer 2007 there will be support for ten students). Participants work in teams of two under my supervision and in consultation with expert programmers. Admission is competitive and based on motivation, strength of the academic record, and letters of recommendation. As a result of taking part in this program, students become better prepared to enter professional scientific careers. UNCG is classified as a Doctoral/Research-Intensive university and is one of the sixteen branches of the University of North Carolina system.

Received by the editor October 16, 2006.

I have delineated the following five goals or objectives for student activities in this REU site program: A first objective is to introduce undergraduate students to various challenging algorithmic combinatorial problems on partial words related to coding, primitivity testing, and computing periods, through lectures and reading. Two types of research opportunities are emphasized: (1) computer related research, with students writing programs to perform experiments on partial words and to implement algorithms; and (2) combinatorics related research, with students investigating properties on partial words to generate conjectures and to discover algorithms. In addition, students are exposed to the techniques of language theory since this is a natural framework for formalizing and investigating strings and operations on them. While achieving this objective, a second objective of the program is for students to develop superior skills in mathematical writing and oral communication, skills that are essential for conducting research in a variety of scientific disciplines. A third objective of this program is to submit the produced original collaborative research on algorithmic combinatorics on words to leading journals. I am working very closely with the students on various research projects supported by this program. High quality publications involving student co-authors are resulting from this work. Also, students gain experience in communicating mathematics verbally through presentations at national professional meetings and national/international conferences. A fourth objective is for students to gain experience in the use of computers and their interaction in mathematical research. As a result, students establish World Wide Web server interfaces for automated use of the programs related to our combinatorial algorithms. This objective involves extensive computer programming and requires some experience using a programming language such as Java. Although students are selected based on merit after a nationwide search from a broad range of colleges and universities, a fifth objective of the program is to strongly encourage underrepresented groups including minorities, women, and students with disabilities to participate.

b. Student Activities

First, as a research mentor, I provide student participants with theoretical background on words and partial words. I use chapters of Lothaire's three books on combinatorics on words as well as my own book "Algorithmic Combinatorics on Words" (currently under review) as required or suggested reading. Second, students are introduced to the language theory techniques that proved to be useful in my prior investigations on words and that enable them to address open problems on partial words (some examples are discussed below). Most of the procedures are based on similar techniques, as well as techniques related to graph algorithms. Third, participants are introduced to some open problems from algorithmic combinatorics on words which are at the early stages of development and offer many challenging opportunities for future research. I create a booklet on each problem that contains some background material, related papers, etc .... The students are divided into teams of two (some choose to work alone) and allowed to pick one problem on which to focus for the remainder of the REU. Students have access to the computing laboratory where they can experiment on words. This facility (Bryan 330) is reserved from June 1 to July 31 of each summer for the participants in this program. A nearby classroom (Bryan 335) is also reserved and made available to

support this program. This arrangement is convenient since my office is near these facilities. Students also have access to the campus library.

One of the goals of the program is to help students attain a higher level of independence in mathematical research. I provide students with some gentle guidance throughout their research projects. This, I believe, gives them independence as researchers and brings them success in their professional lives. The algorithmic combinatorial results get published in major journals, get implemented, and World Wide Web sites get created. Students, who are expert programmers, help me by conducting tutorials on LaTeX, CSS and XHTML, consulting in programming, maintaining the computing laboratory, etc ....

An especially important component of this program is afforded by numerous opportunities to meet with and learn from national/international guest lecturers. These speakers not only discuss their own research, but also discuss topics related to their development as mathematicians or computer scientists. The students (and myself) profit from these seminars and fruitful discussions. In the summer of 2005, I invited in particular Paul Duvall who talked about his expertise with the National Security Agency, and in the summer of 2006, I invited the well-known researcher Jeffrey Shallit from the University of Waterloo who talked about the Thue-Morse sequence. Students also have the opportunity to participate in a wide range of activities outside of the computing laboratory and classroom. At the end of each summer, we have a farewell picnic at Hanging Rock, in North Carolina.

c. Research Projects

I am now giving specific examples of major research findings by some of the student participants from Summers 2005 and 2006:

It is well known that some of the most basic properties of words, like the commutativity ($xy = yx$) and the conjugacy ($xz = zy$), can be expressed as solutions of word equations. An important problem is to decide whether or not a given equation on words has a solution. For instance, the equation $x^m y^n = z^p$ has only periodic solutions in a free monoid, that is, if $x^m y^n = z^p$ holds with integers $m, n, p \geq 2$, then there exists a word w such that x, y, z are powers of w. This result, which received a lot of attention, was first proved by Lyndon and Schützenberger for free groups [**13**]. Dakota Blair and Rebeca Lewis, two REU students from Summer 2005, investigated equations on partial words. When we speak about them, we replace the notion of equality ($=$) with compatibility ($\uparrow$). Among other equations, Dakota, Rebeca and I solved $xy \uparrow yx$, $xz \uparrow zy$, and special cases of $x^m y^n \uparrow z^p$ for integers $m, n, p \geq 2$ [**3**].

Nathan Wetzler, an REU student from Summer 2005, considered one of the most fundamental results on periodicity of words, namely the critical factorization theorem [**9**]. More specifically, given a word w and nonempty words u, v satisfying $w = uv$, the *minimal local period* associated to the factorization (u, v) is the length of the shortest square at position $|u| - 1$. The critical factorization theorem shows that for any word, there is always a factorization whose minimal local period is equal to the minimal period (or global period) of the word. Crochemore and Perrin [**10**] presented a linear time algorithm (in the length of the word) that finds a critical factorization from the computation of the maximal suffixes of the word with respect to two total orderings on words: the lexicographic ordering related to a fixed total ordering on the alphabet, and the lexicographic ordering obtained by reversing the

order of letters in the alphabet. By refining Crochemore and Perrin's algorithm, Nathan and I gave a version of the critical factorization theorem for partial words [**8**]. Our proof provides an efficient algorithm which computes a critical factorization when one exists. Our results extend those of Blanchet-Sadri and Duncan for partial words with one hole [**4**].

Joshua Gafni and Kevin Wilson, two REU students from Summer 2006, introduced the notions of binary and ternary correlations, which are binary and ternary vectors indicating the periods and weak periods of partial words. Extending a result of Guibas and Odlyzko [**12**], Joshua, Kevin and I characterized precisely which of these vectors represent the (weak) period sets of partial words and proved that all valid correlations may be taken over the binary alphabet. We showed that the sets of all such vectors of a given length form distributive lattices under inclusion. We also showed that there is a well defined minimal set of generators for any binary correlation of length n and demonstrated that these generating sets are the primitive subsets of $\{1, 2, ..., n-1\}$. Finally, we investigated the number of correlations of length n and the number of partial words sharing a given correlation [**5**].

Fine and Wilf's well-known theorem states that any word having periods p and q and length at least $p + q - \gcd(p, q)$ also has $\gcd(p, q)$, the greatest common divisor of p and q, as a period [**11**]. Moreover, the length $p+q-\gcd(p, q)$ is critical since counterexamples can be provided for shorter words. This result has since been extended to partial words. More precisely, any partial word u with H holes having weak periods p, q and length at least the so-denoted $l_H(p, q)$ also has period $\gcd(p, q)$ provided u is not $(H,(p, q))$-special. This extension was done for one hole by Berstel and Boasson [**1**] (where the class of $(1,(p, q))$-special partial words is empty), for two or three holes by Blanchet-Sadri and Hegstrom [**6**], and for an arbitrary number of holes by Blanchet-Sadri [**2**]. Taktin Oey and Tim Rankin, two REU students from Summer 2006, further extended these results, allowing an arbitrary number of weak periods [**7**]. In addition to speciality, the concepts of intractable period sets and interference between periods play a role.

d. Student Recruitment

We are attracting outstanding students to participate in this high quality REU program. Eight undergraduate students who have a solid background in mathematical sciences are carefully selected to participate in this eigth-week summer research program in algorithmic combinatorics on words. The students are chosen from a national applicant pool that has exceeded 100 applicants and comes from a broad range of colleges and universities that grant at least the bachelor's degree in Mathematics and/or Computer Science. Institutions targeted include women's colleges and predominantly minority institutions since one objective of this program is to increase participation of women and minorities in science. Outstanding students from universities such as Harvard, Michigan at Ann Arbor, Pennsylvania, Cornell, etc ... have participated so far. I have valuable experience in selecting applicants as I have served on the panel for Mathematical Sciences of the NSF Graduate Research Fellowships Program in 2002, 2003, 2004, 2005, and 2006. A World Wide Web site (continually evolving) has been designed at `www.uncg.edu/mat/reu` for this REU program that contains the program announcement, my publications with undergraduates, application materials, etc .... I was invited to give a talk on my REU site (as well as a plenary talk on partial words) at the *SCRA 2006-FIM XIII*

13th International Conference of the Forum for Interdisciplinary Mathematics on Interdisciplinary Mathematical and Statistical Techniques that was held in Tomar, Portugal from September 1 to September 4, 2006.

NSF support through this REU program is intended for highly motivated students whose undergraduate study is in Mathematics and/or Computer Science. The ideal candidate for this program will have taken a wide variety of upper-level mathematics and/or computer science courses including some of the following: Discrete Mathematics, Combinatorics, Algorithms, Theoretical Computer Science, and Programming. A distinguished academic record and indication of research interest or potential are essential.

e. Program's Success

Student participants are interviewed regularly in order to determine the effectiveness of the REU approach and to provide input for improvements in the program. For example, our program's success with Objectives 2, 3 and 4 are accomplished through the following:

Objective 2: Develop superior skills in mathematical writing and oral communication. These are important skills that are useful for conducting scientific research and that students should gain through some reading of journal papers, etc .... We evaluate Objective 2 in several ways: I provide students the opportunity to critique their peers' work. Also, in the middle and at the end of their REU summer's research, students formally present their results in a paper (typed in LaTeX) and in an oral powerpoint presentation.

Objective 3: Submit the results of our investigations to appropriate peer-reviewed mathematics/computer science journals. The projects I offer are sufficiently sophisticated to permit publication in respected journals, and this has been a typical outcome of prior research projects with my undergraduate students. The NSF supported research work with my REU students is leading to numerous papers submitted to leading journals such as *Journal of Combinatorial Theory, Series A*, *Discrete Applied Mathematics* and *Theoretical Computer Science*.

We extend the research experience beyond the eight-week at UNCG by motivating the students to participate in a national professional meeting or a national or international conference. In the falls of 2005 and 2006, several students attended (or will attend) the *Annual Symposium on Foundations of Computer Science* in Pittsburgh, PA and Berkeley, CA respectively. I also encourage the students to present their work at international conferences. The following students have the distinction of having presented their research at such conferences:

- Dakota Blair, Equations on partial words, *MFCS 2006 31st International Conference on Mathematical Foundations of Computer Science*, Stará Lesná, Slovakia, August 28–September 1, 2006.
- Joel Dodge, Counting unbordered partial words, *SCRA 2006-FIM XIII 13th International Conference of the Forum for Interdisciplinary Mathematics on Interdisciplinary Mathematical and Statistical Techniques*, Tomar, Portugal, September 1–4, 2006 (Session on Undergraduate Research in Interdisciplinary Mathematics).
- Nathan Wetzler, Partial words and the critical factorization theorem revisited, *SCRA 2006-FIM XIII* (Session on Semigroups and Languages).

We keep records of participant co-authored publications and presentations arising from this program as a means of evaluating our success in meeting our third objective.

Objective 4: Establish World Wide Web server interfaces for automated use of the programs. Specific research products that include software (or netware) have resulted such as the World Wide Web sites at

`www.uncg.edu/mat/border`
`www.uncg.edu/mat/research/cft2`
`www.uncg.edu/mat/research/correlations`
`www.uncg.edu/mat/research/equations`
`www.uncg.edu/mat/research/finewilf`
`www.uncg.edu/mat/research/finewilf2`
`www.uncg.edu/mat/research/finewilf3`
`www.uncg.edu/mat/research/unavoidablesets`

that relate to algorithmic combinatorics on words from work with my participants from Summers 2005 and 2006.

Our REU participants from Summers 2005 and 2006 are already being accepted into prestigious Ph.D. programs in Mathematics.

References

[1] J. Berstel and L. Boasson, Partial words and a theorem of Fine and Wilf, *Theoretical Computer Science* **218** (1999) 135–141.

[2] F. Blanchet-Sadri, Periodicity on partial words, *Computers and Mathematics with Applications* **47** (2004) 71–82.

[3] F. Blanchet-Sadri, D. Dakota Blair and Rebeca V. Lewis, Equations on partial words, in R. Královic and P. Urzyczyn (Eds.), *MFCS 2006, 31st International Symposium on Mathematical Foundations of Computer Science, August 28–September 1, 2006, Stará Lesná, Slovakia*, Lecture Notes in Computer Science, Vol. 4162 (Springer-Verlag, Berlin, Heidelberg, 2006) 167–178 (`www.uncg.edu/mat/research/equations`).

[4] F. Blanchet-Sadri and S. Duncan, Partial words and the Critical Factorization Theorem, *Journal of Combinatorial Theory, Series A* **109** (2005) 221–245 (`www.uncg.edu/mat/cft`).

[5] F. Blanchet-Sadri, Joshua Gafni and Kevin Wilson, Correlations of partial words, (`www.uncg.edu/mat/research/correlations`).

[6] F. Blanchet-Sadri and Robert A. Hegstrom, Partial words and a theorem of Fine and Wilf revisited, *Theoretical Computer Science* **270** (2002) 401–419.

[7] F. Blanchet-Sadri, Taktin Oey and Tim Rankin, A generalization of Fine and Wilf's theorem for an arbitrary number of weak periods, (`www.uncg.edu/mat/research/finewilf2`).

[8] F. Blanchet-Sadri and Nathan D. Wetzler, Partial words and the critical factorization theorem revisited, (`www.uncg.edu/mat/research/cft2`).

[9] Y. Césari and M. Vincent, Une caractérisation des mots périodiques, *C.R. Acad. Sci. Paris* **268** (1978) 1175–1177.

[10] M. Crochemore and D. Perrin, Two-way string matching, *Journal of the ACM* **38** (1991) 651–675.

[11] N.J. Fine and H.S. Wilf, Uniqueness theorems for periodic functions, *Proceedings of the American Mathematical Society* **16** (1965) 109–114.

[12] L.J. Guibas and A.M. Odlyzko, Periods in strings, *Journal of Combinatorial Theory, Series A* **30** (1981) 19–42.

[13] R.C. Lyndon and M.P. Schützenberger, The equation $a^m = b^n c^p$ in a free group, *Michigan Math. J.* **9** (1962) 289–298.

University of North Carolina, P.O. Box 26170, Greensboro, NC 27402-6170
E-mail address: `blanchet@uncg.edu`