String Theory

Edward Witten
Institute For Advanced Study, Princeton NJ 08540 USA
(Dated: October, 2001)

In the first of these two lectures, I survey how string theory has developed and explain why I think it is exciting. In the second lecture, I recall the arguments for grand unification of the forces of nature and the reasons for suspecting that supersymmetry is relevant to nature at the TeV energy scale.

I. WHY STRINGS?

In this first lecture, I will tell a little bit of how string theory fits into other ideas in physics. Tomorrow, I will try to recall some general ideas about grand unification which are natural without string theory but also fit naturally in the string context.

String theory began with the "dual resonance model" of the late 1960's. In hadronic physics, many resonances were seen, and they were relatively narrow even though the strong interactions are "strong." Narrowness of a resonance means that the interaction responsible for the decay is relatively weak, so this came as a surprise.

Usually, we can use tree level Feynman diagrams to describe scattering by narrow resonances. For example, consider four body scattering of identical spinless particles. The amplitude due to resonances of spin zero and mass M_n, and coupling g_n is

$$-i\sum_n g_n^2\left(\frac{1}{s-M_n^2}+\frac{1}{t-M_n^2}+\frac{1}{u-M_n^2}\right), \tag{1}$$

where I have introduced the usual Mandelstam variables $s=(p_1+p_2)^2$, $t=(p_1-p_4)^2$, $u=(p_1-p_3)^2$, p_i being the momentum four-vectors of the four particles. (The assumption that the resonances all have spin zero is unrealistic for hadronic physics – or string theory. A more realistic version of (1) would include Legendre polynomials of s,t,u in the numerator to reflect the spins of the resonances.) The three terms in (1) come from diagrams with the n^{th} resonance appearing as a pole in the s-channel, t-channel, and u-channel, respectively. With particles 1 and 2 understood as initial particles and the others as final particles, resonant scattering corresponds to the s channel diagram, while the others, which are more similar to potential scattering than to resonant scattering, are needed for crossing symmetry.

In general, if the center of mass energy is near a resonance of mass M, resonant scattering dominates, the amplitude being approximately $-ig^2/(s-M^2)$. (For a more realistic account one must take account of the width of the resonance, which is important near the pole.) But at energies far from any resonance, resonant scattering does not dominate. For example, if there are only finitely many resonances, and we take s very large at fixed t (corresponding to high energy scattering at small angles) then the resonant scattering diagrams all vanish and the dominant scattering comes from the t-channel exchange.

In hadron physics, there were so many resonances that it seemed one was always near a resonance. This led to a radical proposal called "duality,"[1] which asserted that resonant scattering is the whole answer.

As we have just seen, this is impossible if there are only finitely many resonances. What if there are infinitely many of them? In an intense period of work from about 1968-73, an amazing way to achieve duality was found. A meson was interpreted (figure 1) as a string with "charges" at the endpoints that carry the flavor quantum numbers. The various meson resonances were interpreted as vibrational states of the string. An ordinary tree level Feynman diagram, as sketched in figure 2, was reinterpreted as a string diagram. But in the string picture, the s and t channel diagrams are in fact different limits of the same thing. Thus was duality achieved.

Just around the same time that the string picture was formed, asymptotic freedom was discovered and made possible, in QCD, a more precise and successful theory of the strong interactions. Yet there has always been a striking analogy between QCD and string theory. If the hypothesis of quark confinement in QCD is true in its usual form, than a widely separated quark and antiquark are joined by a "color flux tube." This has an obvious analogy to the notion of a meson as a string with charges at its ends, as assumed in string theory. Explaining

FIG. 1: A meson represented as a string with charges at the ends.

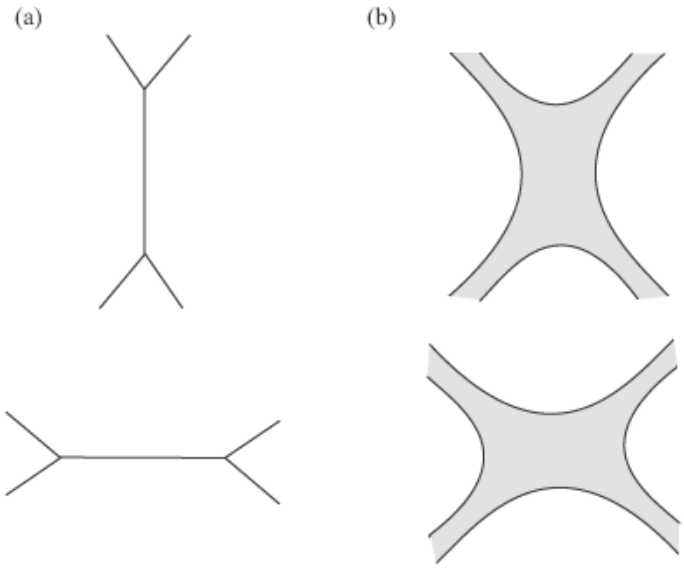

FIG. 2: The $s-$ and t-channel diagrams in field theory (left) and their counterparts for mesons understood as open strings (right). To go from point particles to strings, each line in a Feynman diagram is replaced by a thin strip representing propagation of a string. The strings join smoothly to describe interaction events. Crucially, in the string case, the s- and t-channel diagrams are actually equivalent as the two diagrams on the right can be deformed into one another.

this analogy would mean understanding quark confinement. This would be quite a nice achievement, since it is a longstanding sore point in theoretical physics that despite real experiments and computer simulations supporting the quark confinement hypothesis and despite a lot of ingenious work explaining qualitative criteria for quark confinement and why this notion is natural, there is no convincing, pencil and paper demonstration of quark confinement in QCD.

In 1974, 't Hooft made a brilliant proposal to explain the analogy between string theory and QCD and thereby solve the riddle of quark confinement (and of computing hadron masses): he proposed that QCD, generalized to have N colors, is equivalent to a string theory with $1/N$ as the coupling constant. The idea was motivated by the structure of Feynman diagrams; 't Hooft determined what diagrams dominate for large N and showed that the result had a striking analogy with the structure of string perturbation theory.

A quarter century later, we still do not know if 't Hooft was right. His suggestion, however, remains the most significant approach to the unsolved aspects of QCD. It has stimulated a lot of interesting work, ranging from soluble models of two-dimensional quantum gravity to striking mathematical discoveries about the moduli space of Riemann surfaces to new discoveries about black holes and quantum gravity. In the last few years, using Maldacena's duality between quantum gravity in Anti de Sitter space and conformal field theory on the boundary of spacetime, we have learned that 't Hooft's idea is correct for some four-dimensional gauge theories, but we still do not know about QCD.

Going back to our main story, the simple replacement of the ordinary Feynman diagram shown in figure 2 with its stringy counterpart has numerous consequences, many of which we are still grappling with. In describing them, I will not try to follow a historical order.

(1) The first point is that the infinities of quantum field theory disappear when we go to string theory. This comes from a qualitative difference between a Feynman diagram and its stringy cousin. In a Feynman diagram, there are interaction vertices at which one line splits into two or two recombine into one (as well as higher order vertices with more lines involved). This occurs at a definite spacetime event (whose location one integrates over in constructing the Feynman amplitude). Examples are seen in the ordinary Feynman diagrams on the left of figure 2. In the corresponding stringy diagram, shown on the right of figure 2, there are no distinguished interaction events; globally an interaction (breaking and rejoining of strings) did occur, but there is no way for different observers to agree on just when this happened. Now in general, infinities in quantum field theory arise in loop diagrams when too many interaction vertices are placed at the same spacetime point; there is no analog of this in string theory as there are no interaction vertices, and (as one learns from a much more detailed analysis) the stringy diagrams lack the ultraviolet divergences that arise in ordinary quantum field theory.

(2) There are many possible quantum field theories, most of them with adjustable dimensionless parameter

such as the fine structure constant or the ratio of the electron and muon masses. This is so because, within certain limits (associated with Lorentz invariance, gauge invariance, and so on), one can select more or less "any" set of particles, and any set of interaction vertices, and construct a theory, at least formally. (I say at least formally because at this level of generality one is quite likely to run into unrenormalizable ultraviolet divergences.) Even after imposing all the known physical principles and restrictions, one can still construct quite a large (infinite) assortment of theories. In string theory, one cannot pick the particles arbitrarily – as they are all states of the same string – and one has no freedom about the interactions of the particles since any small piece of a stringy diagram like that in figure 2 looks like any other small piece.[2] (This contrasts with the Feynman diagram where a small piece containing an interaction vertex looks different than a small piece not containing such a vertex.) So in string theory, once one has decided what the string is, the interactions are determined.

When elaborated more fully, this leads to an absence of adjustable dimensionless parameters (such as the fine structure constant) in the fundamental equations of the theory. Such parameters do enter, however, in the choice of a classical solution or (according to our best understanding today) a quantum vacuum.

The absence of dimensionless parameters like the fine structure constant in the basic equations is both good and bad. The good side is that in principle, if we could one day understand the quantum dynamics and solve for the vacuum, we might be able to calculate the dimensionless parameters that we observe in nature. The bad side is that a theory without dimensionless parameters is hard to understand! (Four dimensional non-abelian gauge theory furnishes a fine and frustrating example – that is why quark confinement is not so well understood.)

(3) For the same reason – once the string is picked, the interactions follow uniquely – there are very few string theories. Their construction is rather subtle and their consistency hangs together by a rather delicate thread. In constructing the consistent string theories (roughly from 1969-84), physicists met a succession of surprises:

(A) The idea of supersymmetry emerged largely because (as first seen on the worldsheet in the Ramond-Neveu-Schwarz model, around 1970), it is needed to make string theory work (to include fermions and avoid tachyons). This model helped motivate the extension to spacetime supersymmetry by Wess and Zumino in 1974, after which spacetime supersymmetry was found in the string by Gliozzi, Olive, and Scherk a couple of years later.

Supersymmetry remains the main new concept of experimental significance that has emerged from string theory – but thirty years later, we still do not know whether it is right! (I will come back to this tomorrow.) I guess that this is one of the reasons for our meeting here in Snowmass.

(B) In conventional quantum field theory, gravity (that is, general relativity) is apparently impossible – because of ultraviolet divergences – but string theory forces gravity upon us! That is, all of the string theories contain a massless spin two field (arising as a vibrational state of the string) with interactions like those of general relativity. At least, this experiment has been done!

Reconciling quantum mechanics with general relativity is an old problem, and (at least in my opinion) string theory has been the only significant idea about it. The emergence of quantum gravity from strings has been the single most important motivation for the intensive study of string theory in the last quarter century.

I cannot really explain today the assertion that string theory automatically generates quantum gravity, but I can extract from what I have already said two clues suggesting that string theory entails a modification in our concepts of spacetime, as a quantum gravity theory would be expected to do. (i) The fact that the string diagram has no distinguished interaction vertex is a hint that the notion of a distinct spacetime event has been lost, suggesting that spacetime has become fuzzy (as happened to classical phase space in the light of quantum uncertainty). This naive statement has been at the root of many fascinating computations of non-classical phenomena that occur in string theory. (ii) Supersymmetry itself is a modification of the structure of spacetime, changing the symmetry group of special relativity and incorporating infinitesimal quantum variables in the description of spacetime.

Supersymmetry is far from the whole story about how the concepts of spacetime are modified in string theory – in fact, to date, we only understand bits and pieces of this story. But supersymmetry is an important piece, and it is the piece that we are by far most likely to be able to directly probe experimentally.

The other main consequence of the scarceness and subtlety of string theories has been to make it hard to determine how well string theory can be matched to the real world. For example, by the early 1980's, there were in Type I and Type II superstring theory consistent theories of quantum gravity unified with gauge fields and charged matter. But it was impossible to make a realistic model of particle physics – above all because it was not possible to incorporate parity violation in weak interactions. New discoveries of anomaly cancellation (by Green and Schwarz) and the discovery of the heterotic string (by Gross, Harvey, Martinec, and Rohm) solved this problem, and the models of particle physics became more realistic.

How realistic are they? In general, properties that do not depend on supersymmetry breaking (gauge groups, quark and lepton quantum numbers) come out elegantly, but there is not a convincing mechanism of supersymmetry breaking. The problem is not to break supersymmetry, but to get a sufficiently stable vacuum *without a*

large cosmological constant after supersymmetry breaking.

How many string theories are there? The traditional answer of about 15 years ago was that there were five string theories, differing by very general properties of the strings. In Type I, the strings are open or closed and are unoriented and insulating; the open strings have charges at their endpoints. This is the model that is most closely analogous to QCD. Type IIA and Type IIB superstrings are closed, oriented, and insulating. Finally, heterotic $E_8 \times E_8$ and $SO(32)$ superstrings are closed, oriented superconductors.

Having five string theories is a great advance compared to the infinity of possible quantum field theories. But it leaves the obvious question, "'If one of these theories describes our Universe, who lives in the other four worlds?"

Unlike many of the deep questions in the subject, this one was actually answered in the 1990's: the string theories are different in the classical limit, but quantum mechanically they are the same. In fact, the different string theories are now understood as different limiting cases of a single, richer theory known provisionally (pending a proper understanding that might lead to a natural name) as M-theory.

To unify the five string theories, we have had to understand "duality" symmetries between different descriptions – symmetries that are only valid quantum mechanically. A central role for symmetries whose existence depends on quantum mechanics may be the start of a new understanding of quantum mechanics.

In building up the duality story, we have had to understand a variety of new ingredients, one of the foremost of which is the D-brane. A D-brane is a mini-black hole on which strings can end. They have the fascinating property that the positions of a system of N identical D-branes are described by $N \times N$ *matrices*. When the matrices commute, their simultaneous eigenvalues are the D-brane positions in the classical sense; in general, they do not commute, and spacetime acquires a fuzziness that is reminiscent of the fuzziness of classical phase space in the light of the familiar noncommutativity $[p, x] = -i\hbar$. D-brane matrices have not yet led to a general understanding of what is behind string theory, but they must constitute a clue.

One can, at least in a thought experiment, make a black hole from anything one wants, tables, chairs, or D-branes. If one uses D-branes, one gets a model of a black hole based on $N \times N$ matrices for some large N. The physics of this is $SU(N)$ gauge theory. This link between black holes and gauge theories has been used to get a better understanding of the Bekenstein-Hawking entropy of a black hole – and of quark confinement in gauge theories.

To conclude, I suppose that these are the main reasons to think that string theory is on the right track:

(I) It neatly makes sense of quantum gravity, about which we have had no other significant idea.

(II) It seems to know about all the best ideas in physics – like gravity, gauge theory, and supersymmetry – whether previously known or unknown by humans.

(III) The successes of the theory depend on a sequence of amazing discoveries that is just not plausible if the theory is an accident, rather than part of the description of nature.

On the other hand, we are far from understanding how this theory works, or how to apply it to nature in detail. The best hope to get significant new input from experiment is presumably via the exploration of supersymmetry.

II. SEARCHING FOR SUPERSYMMETRY

In today's second lecture, I will make a survey of what seem to many theorists like some reasonable expectations for physics, starting at high energies and moving down to low energies.

As we discussed string theory yesterday, I will today begin just below the Planck scale at the GUT scale. The basic fact here is that GUT's – or grand unified theories that combine the $SU(3)$, $SU(2)$, and $U(1)$ of the standard model in a unified group such as $SU(5)$ – offer an attractive package. This was so in the 1970's when these theories were introduced; and it is even more so today as a result of subsequent experiments.

There are a variety of reasons for this:

(A) The fermion content of the standard model looks messy at first sight. A single generation of quarks and leptons makes up five representations of $SU(3) \times SU(2) \times U(1)$, with peculiar fractional hypercharges (and electric charges). But this structure fits grand unified theories – based on $SU(5)$, $SO(10)$, or E_6 – perfectly. For example, in $SO(10)$, a generation of quarks and leptons, plus a right-handed neutrino, fits neatly into a single irreducible representation. It is hard to believe that this is an accident.

By the way, we might want to go on and find an even larger gauge group such that *three* generations of quarks and leptons could be derived from an irreducible representation. It turns out that it is impossible to do this in four-dimensional gauge theory, unless one is willing to add three generations of "mirror" fermions with $V + A$ weak interactions. (Because their masses violate the electroweak symmetry, the mirror fermions could not be much heavier than a few hundred GeV, so their existence is all but excluded by electroweak tests.) By starting in higher dimensions, as in string theory, one can indeed derive three generations of chiral quarks and leptons

from an irreducible multiplet of the gauge group E_8. This was one of the successes of string-based ideas in the 1980's.

(B) The $SU(3)$, $SU(2)$, and $U(1)$ gauge couplings are very different at low energies, but the familiar renormalization group running shows that the couplings can be unified at a large energy scale M_{GUT} – if they obey a certain relation at low energies. There are several key points here:

(1) M_{GUT} turns out to be close to the Planck mass – a mass scale that we know about independently. If GUT's were completely wrong, the inferred M_{GUT} might have been a "random" mass.

(2) The relation among the three couplings is a brilliant success, with modern precision measurements, if we make one change in the viewpoint of the 1970's and include *supersymmetry*. This is actually one area where experiments of the last twenty years have made GUT's more attractive, since twenty years ago, the tests of this relation were relatively rough (and certainly not precise enough to see the role of supersymmetry). The GUT picture, adjusted for supersymmetry, has survived a great improvement in the accuracy.

(C) Unifying quarks and leptons is dangerous because it can lead to quark-lepton transitions and thence to proton decay at too high a rate. GUT's pass this test because M_{GUT} comes out to be so high. In fact, nonsupersymmetric GUT's predicted proton decay at a rate that turned out to be too high. But supersymmetry makes M_{GUT} bigger and supersymmetric GUT's live.[3]

(D) In the late 1970's, it was observed that GUT's would lead generically to neutrino masses, with the simplest order of magnitude estimate being $m_\nu \sim M_W^2/M_{GUT}$ or around 10^{-2} electron volts. By now we know that this is the right answer (at least for mass differences), and this is the second area in which experiment of the last 20 years has confirmed the expectations from GUT's.

Usually, when new experiments make a theoretical framework more attractive, it means that the theory is on the right track. I tentatively conclude therefore that GUT's – or at least their key elements that lead to these successes (string theories can entail subtle modifications of GUT's that preserve these successes) – are likely to survive, and that theories of TeV physics that ruin grand unification will probably not survive. This includes many fascinating and ingenious models, including most models with large extra dimensions that I am aware of, and also models in which electroweak symmetry breaking occurs "dynamically," without an elementary Higgs boson.

About dynamical electroweak symmetry breaking, some other arguments run in the same direction. Precision electroweak tests certainly have not supported these models. And really these models would work much more nicely if the quark and lepton bare masses were zero. Dynamical electroweak symmetry breaking is really an elegant idea, but its most simple and elegant versions lead to zero bare masses for quarks and leptons.

So I think it is reasonable to expect a model that is consistent with grand unification, that is, a model with a light Higgs boson. In fact for grand unification to work, the Higgs mass should be less than about 200 GeV. Otherwise, in renormalization group running, the quartic Higgs self-coupling blows up at an energy below the GUT scale.

The same conclusion comes from the precision electroweak fits. They also point to a Higgs mass less than about 200 GeV (with a central value that is considerably less than this).

So theory and experiment appear to agree, and I consider a Higgs boson below 200 GeV (and probably much below it – possibly at the 115 GeV mass suggested by the final LEP results) to be a pretty good bet.

The case for a light Higgs is sharper, of course, if we assume supersymmetry. Then we expect a Higgs below about 135 GeV or so. Supersymmetry is really implicitly built into what I have said so far, since the supersymmetric GUT's are the ones that work.

The most convincing indications of supersymmetry to me are coupling unification, which we have already discussed, and the hierarchy problem. The latter is the question of why the electroweak scale is so low compared to the Planck scale. Supersymmetry makes the hierarchy natural; this approach to the hierarchy problem has passed precision electroweak tests while others have not.

Not quite on the same plane as coupling unification and hierarchy, but still striking, is the mass of the top quark. A very heavy top quark was predicted 20 years ago to make electroweak symmetry breaking work out in supersymmetry (given certain assumptions about soft breaking terms). At the time, such a heavy top quark was not a canonical expectation; accelerators were being built with the hope of discovering the top quark at much lower energies. After 20 years, the framework of electroweak symmetry breaking based on supersymmetry and a heavy top quark is still attractive.

Supersymmetry has a downside as well. Supersymmetrization (with soft breaking terms) introduces roughly a hundred new couplings and has the potential to spoil a notable success of the Standard Model, namely the suppression of flavor changing neutral currents and of baryon and lepton number violation. Supersymmetry might have been more obvious in experiment already; some superpartners might have turned up below 100 GeV. Or supersymmetrization of the Standard Model might have placed greater constraints on it leading to more striking theoretical successes. There might have been a more compelling theoretical picture of the TeV superworld, with a more convincing picture of the soft breaking terms (and why they don't lead to observed

flavor changing or baryon or lepton number violating interactions).

This very downside, however, makes the potential discovery and exploration of supersymmetry an even more exciting target for experimentalists. If theorists had more satisfactory answers to all the questions, what experimenters find would be less of a surprise. As it is, whatever pattern of soft breaking terms is found will be a surprise (at least to me), with the potential to have a big influence on theory. Are squarks really approximately degenerate, as assumed in some models and derived in others, and if so how large are the deviations from this degeneracy? Are gluino masses really "unified" and if so, again, what is the nature of the deviations? The answers to questions such as these will exclude most models of what underlies TeV scale supersymmetry while perhaps lending support to some. Our best hope to get experimental clues about high energy physics going way beyond what we have now is to explore the superworld.

This work was supported in part by NSF Grant PHY-0070928.

---

[1] Note that a different though fortuitously somewhat analogous usage of this term is more common nowadays. In the current usage, a "duality" is a non-classical equivalence of quantum or string theories. We will come to this usage later.

[2] More exactly, we can distinguish small pieces in the diagram precisely by whether or not they contain part of the boundary. But there is no analog of asking whether a small piece of the diagram contains an interaction vertex.

[3] There is an important and not convincingly resolved subtlety here, which is that supersymmetric GUT's can have serious problems with proton decay induced by operators of dimension five or less. Their contributions are model-dependent. Supersymmetry makes the model-independent dimension six contribution sufficiently small to be compatible with experiment.